An Infrastructure, Tools and Methodology for Evaluation of Multicultural Name Matching Systems

Keith J. Miller, Mark Arehart, Catherine Ball, John Polk, Alan Rubenstein, Ken Samuel, Elizabeth Schroeder, Eva Vecchi, Chris Wolf

The MITRE Corporation
7515 Colshire Drive
McLean, VA 22102
USA

{keith, marehart, cnball, jpolk, rubenstein, samuel, eschroeder, evecchi, cwolf}@mitre.org

Abstract

This paper describes a Name Matching Evaluation Laboratory that is a joint effort across multiple projects. The lab houses our evaluation infrastructure as well as multiple name matching engines and customized analytical tools. Included is an explanation of the methodology used by the lab to carry out evaluations. This methodology is based on standard information retrieval evaluation, which requires a carefully-constructed test data set. The paper describes how we created that test data set, including the "ground truth" used to score the systems' performance. Descriptions and snapshots of the lab's various tools are provided, as well as information on how the different tools are used throughout the evaluation process. By using this evaluation process, the lab has been able to identify strengths and weaknesses of different name matching engines. These findings have led the lab to an ongoing investigation into various techniques for combining results from multiple name matching engines to achieve optimal results, as well as into research on the more general problem of identity management and resolution.

1. Introduction

This paper describes a Name Matching Evaluation Laboratory that is a joint effort across multiple projects. The lab houses our evaluation infrastructure as well as multiple name matching engines and customized analytical tools.

2. Infrastructure

At the foundation of our lab's infrastructure is a flexible data model that has been iteratively refined over the course of our project. It contains several layers of abstraction, and enables both the encapsulation of the concepts and the management of the data needed to perform evaluation runs of multiple name matching systems, possibly configured in multiple ways, against varying name data test sets. In addition, it allows us to track relationships between base name records and their linguistic variants, as well as tracking the type of variation. Finally, the data model allows us to manage multiple ground truth versions for our evaluation data, each applicable to a specific use case, and to apply these truth versions to the test runs of the name matching tools, resulting in multi-dimensional evaluations of the tools. Due to size and complexity, the data model is not shown in the paper, but will be available for viewing during the poster session.

3. Methodology

We employ a standard information retrieval evaluation methodology, adapted from those used in evaluation campaigns such as CLEF (Peters and Braschler, 2001) and TREC (Voorhees and Harman, 2000). That is, we measure precision, recall, and F-score on a carefully-constructed test data set. In addition to TREC and CLEF, we draw lessons from the EAGLES/ISLE projects[1]. Specifically, we begin by determining the purpose of the evaluation, and then define a task context in which the system under evaluation will be used. We then develop our test set by collecting name data to model as closely as possible the type and quality of the data that would be found in the task context as it has been defined. Then, after creating a name list and list of name queries to run against that name list, we create adjudication pools by running the queries against the name list, setting the matching thresholds lower than they would be set in actual use. This is done in order to retrieve as close to all matches from the name list as possible, thus enabling a more accurate measure of recall for the various systems. Finally, the items returned in these adjudication pools are judged by human adjudicators. For our purposes, a "good match" is a match that, given the task context, should be nominated for further review by a human reviewer. The ground truth derived from this process is used to evaluate systems at their operational thresholds. A more detailed discussion of the construction of our data sets and our ground truth adjudication process can be found in (Arehart and Miller, 2008).

4. Tools

In addition to the data sets and methodology described above, we have developed several tools to aid in our evaluation of name matching technologies. Some are analytical tools and some aid in the creation or ground

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

truthing of test data. All of these capabilities ride on top of the IML data model, described above.

4.1 IMAC

The Adjudication tool, IMAC, provides a user-friendly web-based environment for name matching adjudicators to create ground truth data sets. IMAC can be installed as a servlet completely separately from the Name Matching Evaluation Lab, along with the name matches that need to be judged. This way, remote adjudicators can participate in the adjudication process as long as they have an internet connection, without having to be connected to the lab's internal network. In the screen shot in Figure 1, we can see that the adjudicator has chosen the three items in the bottom right-hand corner of the screen as good matches for the query "Mhd Ayman Zahabi."

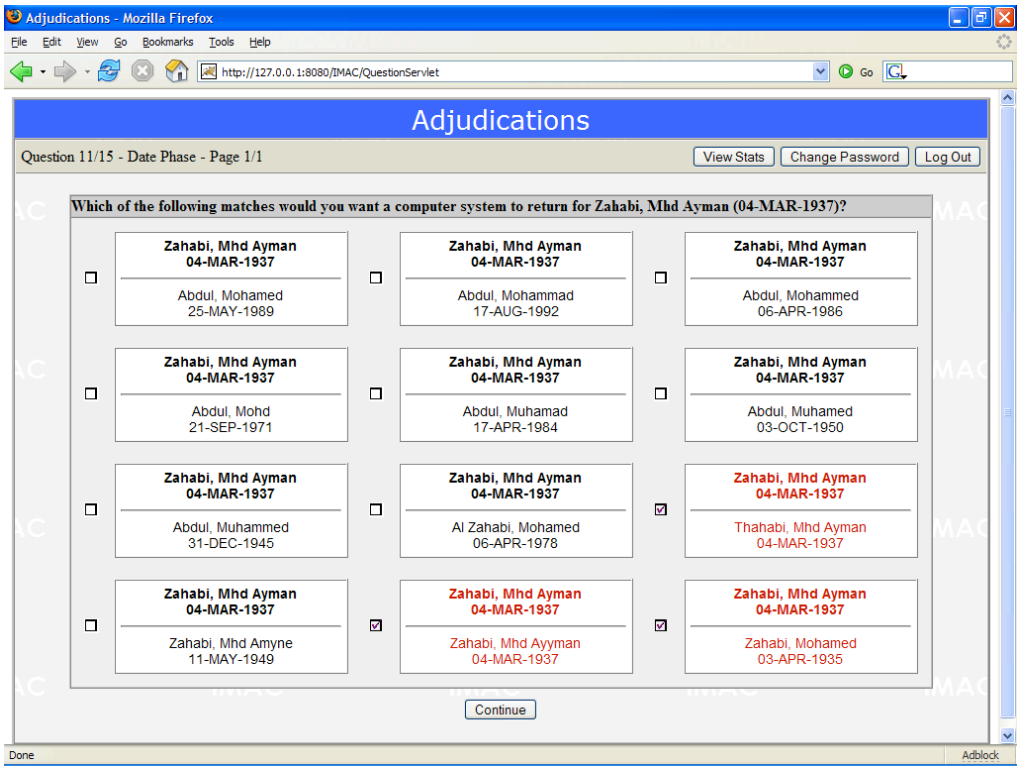

Figure 1: IMAC Adjudication tool

4.2 Ground Truth Compilation

After the raw adjudication data is collected, it must be reconciled into a single version of ground truth. We accomplish this in one of two ways. First, we may have "reconciliation meetings" at which adjudicators who were in disagreement as to the matching status of particular records will discuss their points of view and arrive at a single common judgment. These discussions are guided by a set of adjudication guidelines that were developed at the start of the adjudication effort and that reflect the task context.

Alternatively, we may use an automatic ground truth compilation tool. This tool can be configured to generate a version of ground truth based on the union or intersection of adjudicators' judgments, by favoring the judgments of a particular adjudicator, or by ignoring a particular adjudicator completely. Byproducts of the ground truth compilation procedure include statistics pertaining to interadjudicator agreement.

4.3 R Scripts

Within the R statistical programming environment, we have developed modules for analyzing and graphing the performance of name matching technologies against these ground truth data sets. One sample graph can be seen in Figure 2, which displays the performance of three name matching engines in terms of their precision-recall curve. In this graph, the point of optimal F-score is indicated by an open circle on the P/R curve.

In addition, since the tuning and performance evaluation of the name matching systems depends crucially on the quality of the ground truth data upon which these operations are based, and since the construction of ground truth in the context of name matching is a somewhat subjective process, we have also developed statistical modules to evaluate the quality of our ground truth data, in terms of agreement achieved between expert judges during the adjudication process. We are also currently conducting some research into the relative importance of achieving absolute consensus among adjudicators versus using an adjudication pool that is constructed in such a way as to eliminate bias toward any particular adjudicator. That research is described (Arehart et al, 2008).

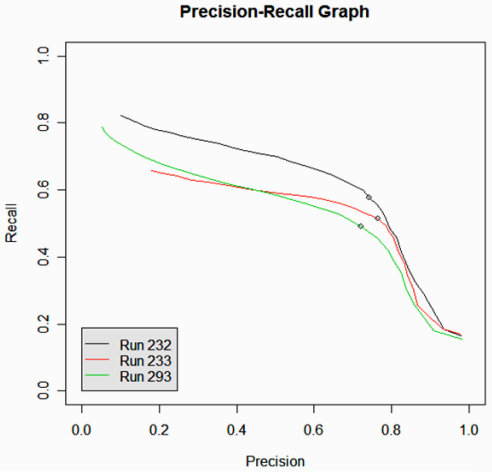

Figure 2: Performance Graphs in R

4.4 MINERVA

In addition to getting the high level view of name matching effectiveness mentioned above, we have also developed tools that enable us to take a "deep dive" look at the actual results being produced by each individual name matching system, and to easily compare results across systems – or of different settings of the same system – at a low level of granularity. MINERVA highlights results using selectable truth, based on human adjudication, to quickly distinguish the desirable from undesirable results. Figure 3 shows a snapshot of a MINERVA session with true positives in green, false positives in red and false negatives (given a certain threshold: 76 in this case) in red text, shaded gray.

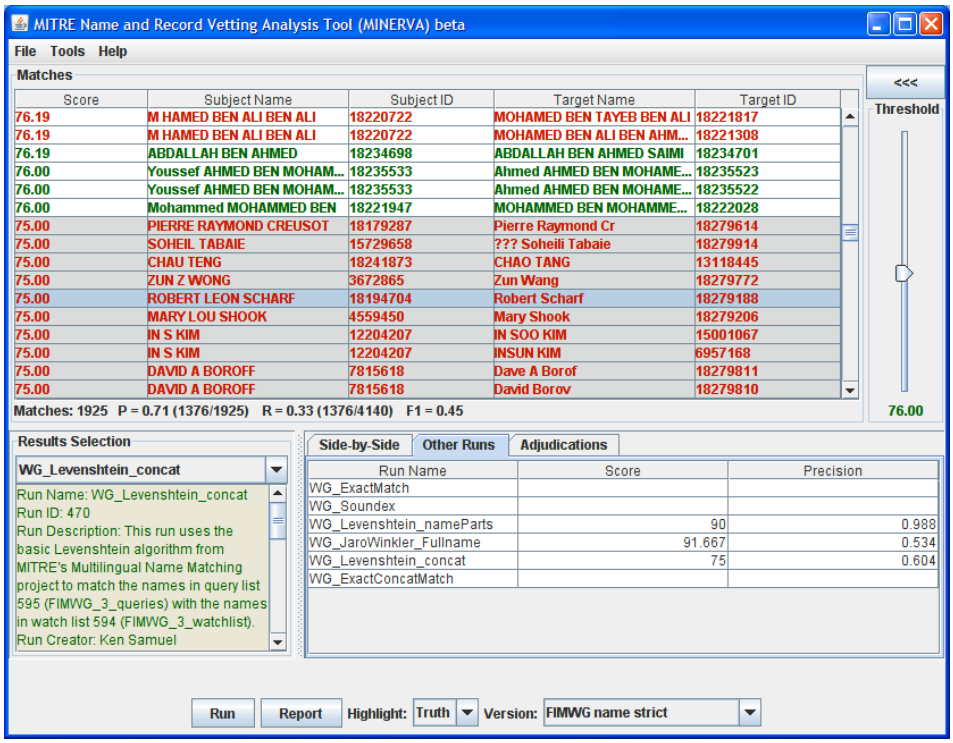

Figure 3: MINERVA – a detailed look at name matching results

Using MINERVA, results from multiple name matching systems can be combined into a "virtual cocktail" of name matching systems that may produce better results than any of the contributing systems alone. The interface for this is shown in Figure 4, where the F-Score for the "cocktail" (0.6054) surpasses that of either system that contributed to the cocktail (0.5946 and 0.4503).

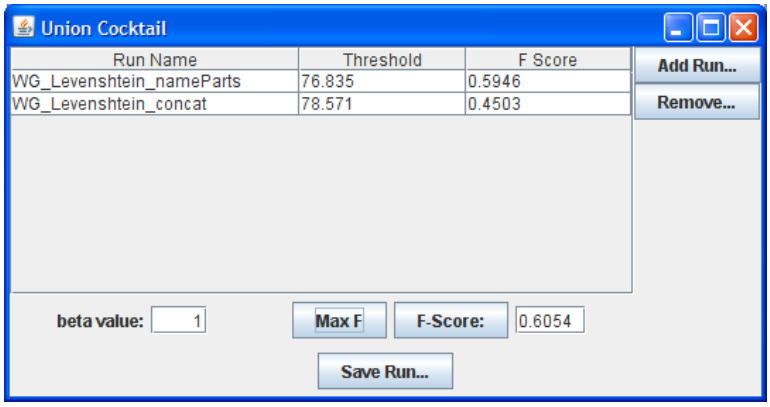

Figure 4: MINERVA – joining results from multiple name matching systems

4.5 GenV

In order to test the limits of the types of linguistic variation a name matching system can handle, we have also developed a tool (called GenV) through which expert users can generate motivated variants of base name data, to be included in test set name lists. During the variant creation process, users of GenV tag each variant with one or more "variant types" according to the taxonomy of name variation, which has been iteratively developed through many hours of interaction with naturally occurring data. The name variation taxonomy can be seen on the right side of the application pictured in Figure 5, with a larger version of the taxonomy in Figure 6.

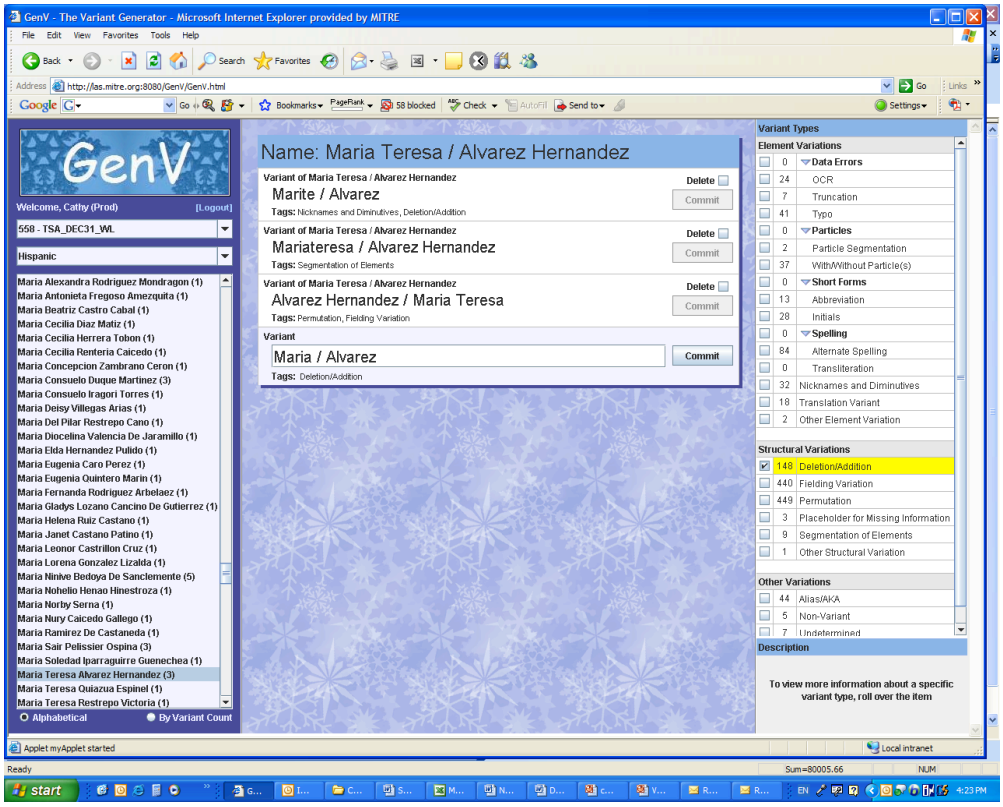

Figure 5: GenV – the Name Variant Generation and Tagging Tool

Element Variations

- Data Errors
  - OCR
  - Truncation
  - Typo
- Particles
  - Particle Segmentation
  - With/Without Particle(s)
- Short Forms
  - Abbreviation
  - Initials
- Spelling
  - Alternate Spelling
  - Transliteration
- Nicknames and Diminutives
- Translation Variant
- Other Element Variation

Structural Variations

- Deletion/Addition
- Fielding Variation
- Permutation
- Placeholder for Missing Information
- Segmentation of Elements
- Other Structural Variation

Other Variations

- Alias/AKA
- Non-variant
- Undetermined

Figure 6: Personal Name Variant Taxonomy

4.6 Picos

Finally, we have developed a utility called Picos, which wraps the evaluation process mentioned earlier into one easy-to-use, step-by-step web application. It can be thought of, in simple terms, as a name matching evaluation wizard. Its intended use is to enable those not familiar with our data model to evaluate their own name matching engines. Picos is automatically populated with the lab's ground truth data, which users can leverage to create scenarios containing query and target lists that match their specific use case. These lists can then be run through the user's name matching algorithm, and the results loaded into Picos. Evaluation statistics for the user's results are then calculated, and can be compared to other algorithms run on the same query and target lists.

5. Conclusion and Future Work

Using the infrastructure, tools, and test data described in this paper, our team has been able to document the relative strengths and weaknesses of each of the name matching tools tested in the lab. Using tools such as MINERVA, the lab's researchers can run various experiments and quantify the accuracy of results, both at a high level that ranks overall system performance and at a low level that reveals which challenges a system handles well and ones it handles poorly.

The IML team is also using these tools to investigate combinations of matching engines that might improve on the performance on any one engine running alone. Improved results are often obtained by using the union of results from two or more engines. This work is still ongoing and is showing promising early results, as described in (Miller and Arehart, 2007).

Future work will focus also on how the robust evaluation platform designed in the IML for name matching might need to be modified and augmented for evaluation of identity matching systems – that is, systems that find matches between records containing multiple types of information in addition to *personal name*. Although we believe that our underlying evaluation methodology will stand up to this challenge, we are certain that performing evaluation of identity resolution tools will involve updating our data model, tools, and possibly our choice of evaluation metrics.

6. References

Arehart, Mark and Miller, Keith J., A Ground Truth Dataset for Matching Culturally Diverse Romanized Person Names, LREC 2008, Marrakesh, Morocco.

Arehart, Mark, Wolf, Chris and Miller, Keith J., Adjudicator Agreement and System Rankings for Person Name Search, LREC 2008, Marrakesh, Morocco.

Miller, Keith J. and Arehart, Mark, Result Aggregation for Knowledge-Intensive Multicultural Name Matching, Language and Technology Conference 2007, Poznan, Poland.

Peters, C., Braschler, M, (2001). "Cross-Language System Evaluation: the CLEF Campaigns", Journal of the American Society for Information Science and Technology, 52(12):1067-1072, 2001.

Voorhees, Ellen and Donna 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.

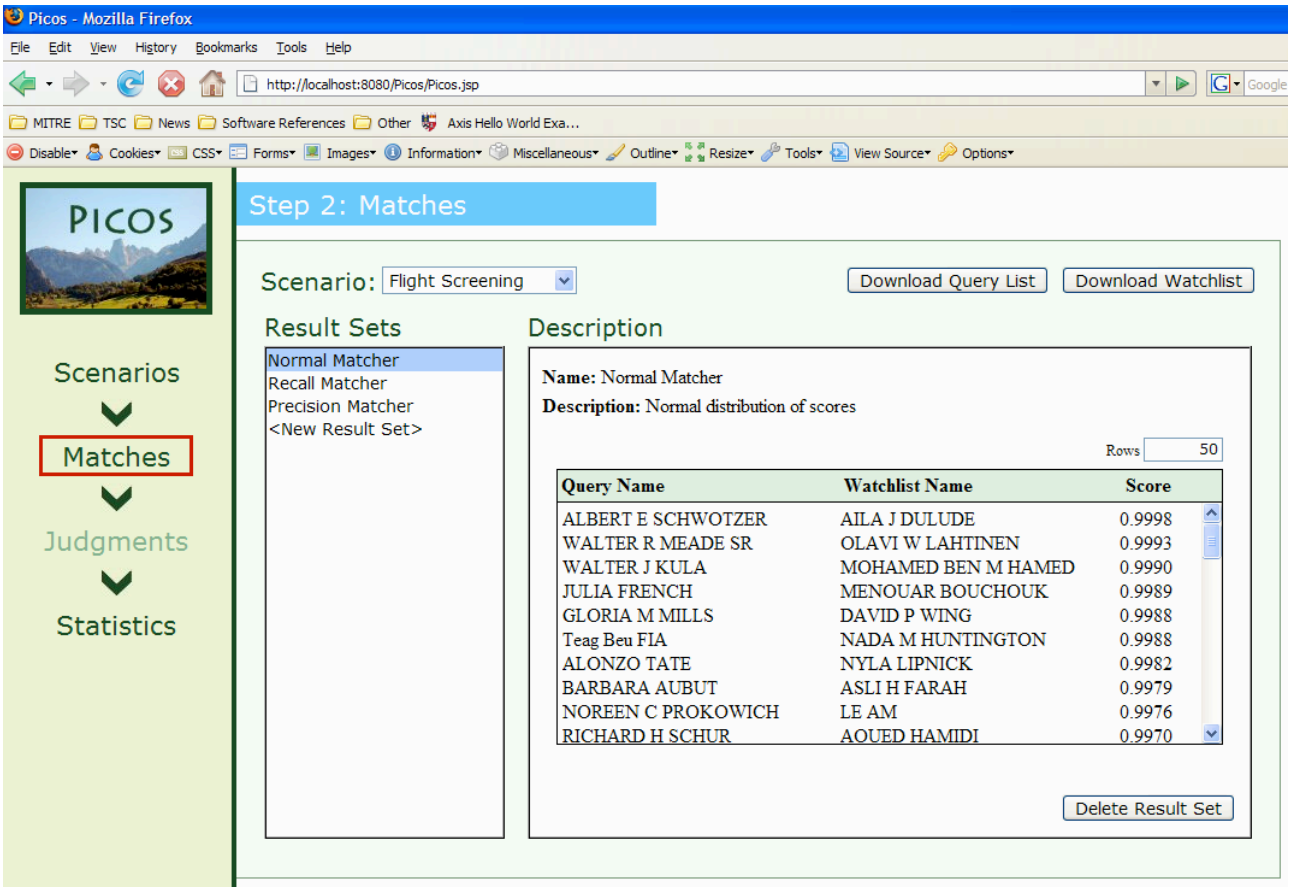

Figure 7: Picos – end-to-end evaluation tool

(Note: name and score data fabricated)