Adjudicator Agreement and System Rankings for Person Name Search

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

Abstract

We have analyzed system rankings for person name search algorithms using a data set for which several versions of ground truth were developed by employing different means of resolving adjudicator conflicts. Thirteen algorithms were ranked by F-score, using bootstrap resampling for significance testing, on a dataset containing 70,000 romanized names from various cultures. We found some disagreement among the four adjudicators, with kappa ranging from 0.57 to 0.78. Truth sets based on a single adjudicator, and on the intersection or union of positive adjudications produced sizeable variability in scoring sensitivity – and to a lesser degree rank order – compared to the consensus truth set. However, results on truth sets constructed by randomly choosing an adjudicator for each item were highly consistent with the consensus. The implication is that an evaluation where one adjudicator has judged each item is nearly as good as a more expensive and labor-intensive one where multiple adjudicators have judged each item and conflicts are resolved through voting.

1. Introduction

1.1. Evaluation and Proper Name Search

Valid and useful evaluation of human language technologies depends crucially on the construction of high-quality ground truth data. Even methodologies employing automated metrics (e.g. BLEU for evaluation of machine translation (Papineni *et al*, 2001)) require this often labor-intensive and expensive step. Thus, a goal of many evaluation methodologies is to minimize the initial cost of developing this ground truth, to maximize its reusability, or both. In this paper, we describe a number of experiments using variously-developed versions of ground truth data for a single data set produced for the evaluation of search engines that specialize in the retrieval of personal names. These experiments indicate that it is possible to achieve significant cost savings in the development of ground truth data for this evaluation purpose while still maintaining high quality.

Romanized proper names exhibit variation in transliteration, database fielding, segmentation, and the number and types of name segments present. The type of variation encountered depends on the linguistic origin of the name and on the way such names are typically represented in Western databases. Identifying plausible high-quality matching name variations is a knowledge-intensive task for a computer system or a human adjudicator.

1.2. Evaluation Use Case

We assume a scenario where the purpose is to determine the relative performance of several name matching algorithms on a dataset of Romanized names of mixed linguistic origin, with performance evaluated using a balanced F-score (F1).

1.3. Benefits of Analyzing Agreement

We see three benefits in assessing the impact of variability among adjudicators on the reliability of system rankings. First, one can determine how much disagreement affects the evaluation results. Second, one can potentially save time and effort. The process of performing adjudications and manually resolving disagreements (e.g. by expert committee) is time-consuming, and by extension, expensive. If system rankings are stable in spite of some disagreement, then some of this effort may be unnecessary. Third, these results provide a comparison with similar studies of other tasks in Information Retrieval.

2. Data Set and Methods

2.1. Test Corpus

We collected names from two publicly available sources. The first is the Death Master File, published by the Social Security Administration, which contains the names of about 77 million deceased holders of social security numbers[1]. Although limited to the United States, the data source is large enough so as to contain names from a variety of linguistic and cultural origins. The second source is the Mémoire des hommes, 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[2]. As such it contains not only French names, but also Southeast Asian and Francophone-transliterated Arabic names.

Using a commercial name culture classification tool, 70,000 names were chosen with a stratified cultural distribution, including Anglo, Arabic, Hispanic, Chinese, Korean, Russian, Southwest Asian (Farsi, Afghani, and

[1] http://www.ntis.gov/products/ssa-dmf.asp. We would like to acknowledge Catherine Ball for identifying this data source.
[2] http://www.memoiredeshommes.sga.defense.gouv.fr/

Pakistani), French, German, Indian, Japanese, and Vietnamese.

Additionally we manually created 1,146 variants on 404 (about 0.6%) of the base records, averaging 2.8 variants per record.

Because it is infeasible to adjudicate the results of matching the entire list against itself, we chose a subset of 700 as queries. The queries come from two groups: the 404 "base" records, and randomly selected records. Of these 700 queries that were used in a larger evaluation, 100 were randomly selected for this study.

2.2. The Adjudication Task

We created adjudication pools by adapting the methodology of the National Institute for Standards and Technology (NIST) Text REtrieval Conference (TREC) (Voorhees and Harman, 2000; Voorhees, 2001). To create adjudication pools, results were aggregated from several open source and commercial tools using lower-than-normal matching thresholds.

To be maximally useful, evaluation should be done with reference to a particular use context. For information retrieval, one consideration is the relative importance of precision and recall, or, put another way, the tolerance for false positives and false negatives. In the use case envisioned for this evaluation, a system presents name search results to a user who then has access to additional identifying attributes to make a decision about an overall identity match. Further, we imagine a scenario in which the cost of a false negative is relatively high. Thus, this user is willing to sift through spurious matches in order to ensure that she does not miss a potentially good identity match.

We therefore developed a set of guidelines using a "loose" truth criterion, by which two names should be considered a match despite variation beyond superficial spelling differences, as long as there is a plausible relationship between the names. The guidelines enumerate several types of name variations that can establish such a relationship, including both segment-level variation (e.g. alternate spellings) and structural variation (e.g. additions, deletions, reorderings). For example, the names in Figure 1, in which the data contained in the surname field is capitalized, would be considered a possible match.[3]

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

Figure 1: Arabic name variation.

Although the adjudicators varied in their level of domain expertise, all had some knowledge of linguistics and had the opportunity to read and discuss the adjudication guidelines.

Four adjudicators completed the pools for the 100 queries used in this study. Exact string matches were excluded, leaving 1712 total common items upon which agreement was assessed.

2.3. Alternate Truth Sets

Systems were scored with the alternate truth sets described in Table 1. Each type represents a different means of resolving multiple adjudications into a single true/false decision.

	Type	Criteria for true match
1	Consensus	Tie or majority vote
2	Union	Judged true by anyone
3	Intersection	Judged true by everyone
4	Single	Judgments from a single adjudicator
5	Random	Randomly choose adjudicator per item

Table 1: Truth sets.

There are as many Single versions of truth as there are adjudicators. There are n^i Random truth sets, where n is the number of adjudicators and i is the number of items judged, of which the other types are special cases. One thousand Random truth sets were generated for analysis.

2.4. Comparing System Rankings

Direct comparison of two rankings is problematic when one accounts for the significance of score differences. Consider a ranking of algorithms A, B, and C where A is not significantly better than B, and B is not significantly better than C, but A is significantly better than C. A ranking of A>B>C implies more significance than is present, and a tied rank for all three obscures the difference between A and C.

We define the evaluation results as a set of evaluation statements about pairs of algorithms, where for any pair A and B there are three possible statements: A>B, B>A, and A=B (with the operator ">" indicating statistically significant difference, and "=" no statistically significant difference). Another way to conceptualize the results is as a partial ordering of systems, where the ordering relation is a statistically significant difference. For an evaluation with n systems, there are $n(n\text{-}1)/2$ evaluation statements, derivable from the combination formula shown in Figure 2.

[3] Because of the structure of Arabic names, 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*.

$$C_k^n = \frac{n!}{k!(n-k)!}$$

Figure 2: Combination.

In our case, where we have evaluated 13 algorithms, this yields 78 statements. We look at the sensitivity of results of a set of statements (the proportion of algorithm pairs showing significant differences), the rate of disagreement between statements derived from two truth sets (the proportion of statements that do not agree), and the proportion of results that are reversed between two truth sets. A reversal means that under one truth set A>B, and in the other, B>A. For purposes of comparison, we take the Consensus truth set as the baseline.

We score systems using F1, the harmonic mean of precision and recall. F1, which is neither a proportion nor a mean of independent observations, is not amenable to traditional statistical tests, so we use bootstrap resampling to test for significance (Efron and Tibshirani, 1993; Bisani and Ney, 2004; Keller *et al*, 2005), with a significance level of 0.05. In our implementation we used the "shift" procedure described in Noreen (1989) and Riezler and Maxwell (2005).

3. Results

3.1. Levels of Agreement

Adjudicator agreement was computed in several ways. Overlap is the number of positive judgments in common divided by the total number of positive judgments. The statistics p+ and p- are the proportions of specific agreement on positives and negatives, respectively (Fleiss, 1981). The formulas for overlap and specific agreement are shown in Figure 3, based on a standard contingency table where cell *a* represents the number of agreements on positives, *d* the number of agreements on negatives, and *b* and *c* the number of disagreements.

overlap	=	a / (a + b + c)
p+	=	2a / (2a + b + c)
p-	=	2d / (2d + b + c)

Figure 3: Agreement formulas.

Figure 4 shows pairwise agreement between adjudicators, labeled A through D, and also includes the commonly used kappa metric (Fleiss, 1981). The lowest agreement is between adjudicators A and B, where kappa is 0.57.

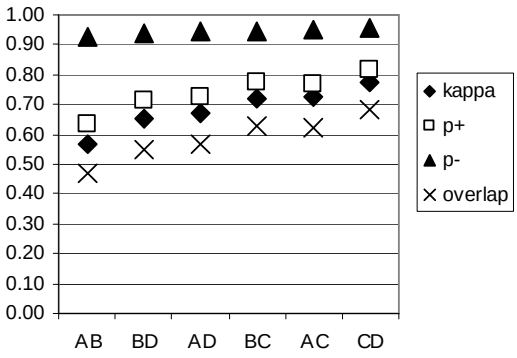

Figure 4: Adjudicator agreement.

Figure 5 shows the base rates of acceptance. Recall that adjudication pools are designed to include many false matches in order to increase the likelihood that they contain all the true matches. Thus the low acceptance rates are expected.

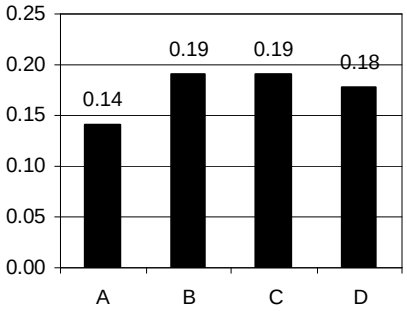

Figure 5: Base rates of acceptance.

3.2. System Rankings

3.2.1. Baseline Versus Random

The baseline (Consensus) truth set yielded an evaluation set with a sensitivity of 0.744, meaning that about three quarters of the pairwise algorithm comparisons showed a significant difference in score. The 1000 Random truth sets had a mean sensitivity of 0.729, with a 0.05 confidence interval of 0.728 to 0.731, which is therefore significantly lower than the sensitivity of the Consensus truth set. The mean level of disagreement between Consensus and Random truth sets was 0.0727 (0.05 confidence interval: 0.0714, 0.0742).

The disagreements are entirely attributable to differences in sensitivity, as there was not a single example where a significant difference in the baseline truth set was reversed in a Random truth set. In other words, using a Random truth set in lieu of the baseline set slightly reduces the ability to detect differences, but one will never predict that A>B if the baseline set predicts B>A, or vice versa. Because significance was computed at the 0.05 level, one expects the baseline level of disagreement to be at least 0.05. Indeed, the expected level of disagreement is higher

because of the simultaneous testing of multiple hypotheses. This leaves 0.0227 or less, depending on how one corrects for multiple hypothesis testing, attributable to the difference in the method for compiling the truth set. Based on these results, there appears to be little practical difference in the results of an evaluation based on Consensus versus Random truth sets.

3.2.2. Baseline Versus Special Cases

Table 2 shows sensitivity and two comparisons to the baseline (Consensus) truth set: proportion of disagreement and proportion of reversed statements.

Truth Set	Sensitivity	Disagreement	Reversal
Consensus	0.744	n/a	n/a
Union	0.782	0.064	0
Intersection	0.538	0.423	0.038
Judge A	0.769	0.051	0
Judge B	0.705	0.038	0
Judge C	0.756	0.115	0
Judge D	0.692	0.179	0

Table 2: Truth Set Comparisons

Compared to the Random truth sets, there are greater differences in sensitivity and more varied levels of disagreement, ranging from a low of 3.8% for judge B to a high of 17.9% for judge D. The Intersection truth set was the only one where a significant difference in the baseline set was reversed. The 3.8% reversal rate represents three pairwise comparisons. All three include a single algorithm whose ranking dropped under the Intersection truth set.

Figure 6 shows selected system F-scores under different truth sets. One notable outcome is that the system that is either first or tied for first in all other rankings performs relatively poorly in the Intersection truth set. As it turns out, that system has the highest recall of any tested, and presumably suffers the most under the stricter match criteria. A more detailed analysis of specific systems' sensitivity (or lack thereof) in performance to different truth sets is a topic for future work.

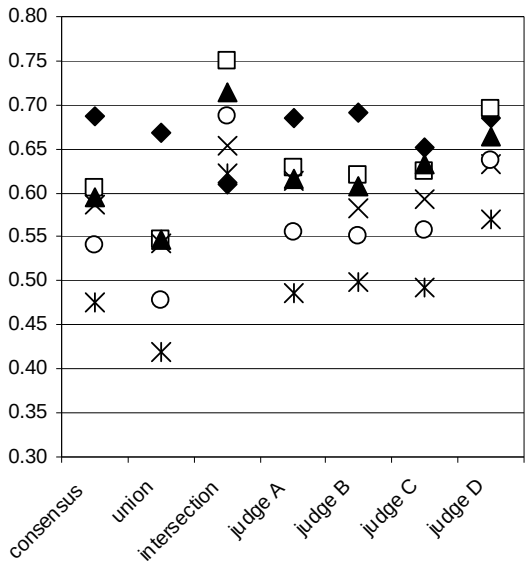

Figure 6: Selected System F Scores.

4. Related and Future Work

4.1. Comparison with TREC Data

Voorhees (2000), in an experiment on TREC data, found IR system rankings to be stable in the face of varying relevance judgments. That study compared system rankings based on mean average precision using Kendall's tau, and also analyzed the probability of swaps (or reversals) of rankings between truth sets. However, it did not apply a test to determine which ranking differences were statistically significant. Had the similarity of rankings not been penalized by non-significant differences, it is likely that the similarities in evaluation outcomes would have been even more robust. Because of the different measures used, the results of our study, though roughly consistent with Voorhees' work, are not directly comparable.

4.2. Breakdown By Culture and Variant Type

This adjudication task, unlike the TREC task, requires specialized linguistic knowledge, only some of which can be adequately covered in the adjudication guidelines. An area for future research is to determine the sources of disagreement. One factor is different perceptions of the similarity threshold distinguishing matches from non-matches. Another factor is different perceptions of what constitutes similarity, which may vary among different name cultures and types of variation.

5. Conclusion

This study has shown that results based on a truth set compiled from the judgments of different adjudicators, each judging a different subset of matches, is highly consistent with results from a truth set representing the group consensus on every item. Although the consensus

truth set is slightly more sensitive to score differences, there are no reversals of results for the 13 systems tested, at least not when using an appropriate significance test. Results are more varied, however, when using the judgments of a single adjudicator or the union or intersection of matches from all adjudicators. Roughly speaking, differences among adjudicators appear to "wash out" in the Random truth sets and therefore approximate the consensus. In contrast, the Union, Intersection, and Single truth sets exhibit more varied characteristics.

References

Bisani, M. and H. Ney. (2004). Bootstrap Estimates for Confidence Intervals in ASR Performance Evaluation. In Proc. *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Montreal, Canada, Volume 1, pp. 409-412.

Efron, B. and R.J. Tibshirani. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*. 2nd ed. New York: Wiley.

Keller, M., S. Bengio, and S.Y. Wong. (2005). Benchmarking Non-Parametric Statistical Tests. *Advances in Neural Information Processing Systems* 18. Vancouver, BC, Canada.

Noreen, E.W. (1989). *Computer Intensive Methods for Testing Hypotheses: An Introduction*. New York: Wiley.

Papineni, K.A., S. Roukos, T. Ward, W.J. Zhu. (2001). *BLEU: a method for automatic evaluation of machine translation.* Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center.

Riezler, S. and J.T. Maxwell. (2005). On Some Pitfalls in Automatic Evaluation and Significance Testing for MT. *Proc. of the ACL-05 Workshop on Intrinsic and Extrinsic Evaluation Methods for MT and Summarization*, Ann Arbor, Michigan.

Voorhees, E.M. (2000). Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. *Information Processing and Management* 36(5), pp. 697-716.

Voorhees, E.M. (2001). The Philosophy of Information Retrieval Evaluation. *Lecture Notes in Computer Science* 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, ed., *The Eighth Text REtrieval Conference* (TREC-8), Gaithersburg, MD, USA.