# International Multicultural Name Matching Competition: Design, Execution, Results, and Lessons Learned

**Keith J. Miller, Elizabeth Schroeder Richerson, Sarah McLeod, James Finley, Aaron Schein**
The MITRE Corporation
7515 Colshire Drive
McLean, VA 22102 USA
keith@mitre.org, eschroeder@mitre.org, smcleod@mitre.org, jtfinley@mitre.org, aschein@mitre.org

**Abstract**

This paper describes different aspects of an open competition to evaluate multicultural name matching software, including the contest design, development of the test data, different phases of the competition, behavior of the participating teams, results of the competition, and lessons learned throughout. The competition, known as The MITRE Challenge™, was informally announced at LREC 2010 and was recently concluded. Contest participants used the competition website (http://mitrechallenge.mitre.org) to download the competition data set and guidelines, upload results, and to view accuracy metrics for each result set submitted. Participants were allowed to submit unlimited result sets, with their top-scoring set determining their overall ranking. The competition website featured a leader board that displayed the top score for each participant, ranked according to the principal contest metric - mean average precision (MAP). MAP and other metrics were calculated in near-real time on a remote server, based on ground truth developed for the competition data set. Additional measures were taken to guard against gaming the competition metric or overfitting to the competition data set.

**Keywords:** multicultural name matching, evaluation, challenge competition

## 1. Introduction

Person name matching is a problem that arises when different versions of a person's name that exist in multiple sources must be identified as being related. A person can have multiple versions of a name for reasons such as using nicknames or reaching life milestones, such as marriage, making the Haj to Mecca, or receiving an advanced degree. Transliteration, translation, or simple data input errors can also lead to variations of the same name. These variations can be problematic, for example, when searching for a patient record at a doctor's office, screening an individual before granting them access to sensitive information, or merging or deduplicating records in a customer database. In these situations, and many others, multicultural person name matching technology is used to find person name records that would otherwise not be located.

In 2011 the MITRE Corporation launched an open competition called The MITRE Challenge™ to evaluate multicultural name matching software. The open format of the Challenge allowed the evaluation of a large range of name matching solutions since no constraints were imposed on the algorithmic approach. Additionally, because virtually anyone could participate, the competition presented an opportunity to identify top-performing solutions that might not have previously received attention, rather than only those from a small set of previously identified solution providers. Participants were allowed to self-select team names for display on the publicly-visible leader boards. By allowing anonymity, the Challenge hoped to attract participants who might be wary of having performance metrics posted publicly.

## 2. Contest Design

The MITRE Challenge was inspired by the Netflix Prize – a competition sponsored by Netflix challenging participants to improve the Netflix movie recommendation algorithm (www.netflixprize.com). Netflix provided participants with training data composed of anonymous user ratings, as well as test data sets, which participants ran through their algorithms and submitted to Netflix for scoring.

Because of the open format of the Netflix Prize virtually anybody could participate, including experienced professionals and curious amateurs. The data-driven nature of the competition allowed participants to develop solutions to achieve a goal, rather than to conform to a set of requirements. These aspects of the Netflix Prize parallel MITRE's goals to identify a broad set of ideas and solutions to various challenges.

### 2.1 Domain

The initial MITRE Challenge focused on the evaluation of multicultural person name matching systems. This domain was chosen for two reasons: First, name matching has a broad range of uses, ranging from the support of screening and credentialing services to disaster relief, benefits distribution and fraud prevention. Second, MITRE has previous experience in evaluating person name matching software, and already had an existing infrastructure for carrying out those evaluations (Miller et al, 2008) upon which the Challenge could be based.

### 2.2 Data

To create the competition data set, MITRE identified various sources from which person names could be

collected. To minimize the time spent collecting names for the data set, MITRE only considered sources in a structured format that was relatively easy to parse. Each source considered contained a significant number of person names, with at least a few thousand names each. Smaller sources, with fewer than several thousand but at least a few hundred person names, were considered if the names in that source represented a particular culture or variation type not found in larger sources. All names collected for the data set were rendered in the Roman alphabet. Additionally, we considered only sources that complied with privacy laws in the United States.

The Challenge data set consisted of two lists, a query list and an index list. Both the query and index lists contained names from a variety of cultures and source languages, though all names were rendered in Roman script. Names displayed a number of different types of variation, due to both the nature of source data sets and the inclusion of hand-created variants of names in the data. Hand-created variants were based on the variation taxonomy described in Miller, et al, 2008.

Of the names in the query and index lists, the length of the full name records ranged from four characters to 69 characters, with zero to 49 characters in the given name, and zero to 39 characters in the surname. Full names ranged from one to 11 segments, where name segments are delimited by a single white space. The shortest given and surnames had zero segments, or were null, while the largest number of given name segments was nine and surname segments was six. Names were presented in the format "given name|surname", though as indicated above either the given name or the surname could be null. As is common when dealing with multicultural data, the given name-surname distinction could vary throughout the data set, and was intentionally not restricted to Western-influenced fielding or structuring, such as "Surname, First name Middle Initial."

Both the query and index lists contained distractor data, or names that were not factored into the team metrics, meaning that only a subset of the possible matching name pairs actually contributed to a participant's score. Participants were not aware of the distinction until the Challenge concluded, and had no practical way of determining which names contributed to the metrics. The data set was designed in this manner so as to prevent teams from manually judging all possible name pairs, or from using other "brute force" tactics to otherwise manipulate the contest.

The query list contained a total of 8,666 names, 266 of which contributed to participant scores, and the index list contained 826,388 total names, 36,069 of which contributed to the metrics. Participants were allowed to return a maximum of 500 pairs of matching names for each name in the query list. There were 1,120 correctly-matching name pairs to be identified in the scored data. The identification and adjudication of these names pairs is addressed in Section 3 in the discussion of the ground truth data set.

## 2.3 Metrics

To provide an accurate ranking of systems, MITRE considered multiple metrics for use in scoring participant submissions. The scoring metric had to meet the following requirements:

- All queries contribute equally to the overall score
- Unreturned true matches negatively impact the overall score

Two metrics commonly used in Information Retrieval evaluations, F-score and Mean Average Precision (MAP), met both requirements. MAP, as the name suggests, uses the average precision of each the queries in the data set to calculate the average precision over all queries in the data set. Average Precision for a given query is calculated by averaging over all possible ranks the percentage of relevant documents returned for a given rank.

F-score is the harmonic mean of precision and recall at a given score threshold, where precision is the percentage of correct results in all results, and recall is the percentage of correct results returned from all of the possible correct results.

To decide between MAP and F-score, we considered additional details regarding the metrics. One detail was whether a metric required match scores in a submitted result set to fall within a pre-specified range, such as from 0 to 100. While such restrictions on match scores were acceptable, metrics with no such restrictions were preferred. Neither MAP nor F-score requires match scores to fall in a pre-specified range. Additionally, we considered whether certain implementations of a scoring algorithm handled tied submission scores in a graceful manner. Again, both the MAP and F-score metrics complied with this consideration.

The team also considered whether the metric accounts for the ranking of matches within a result set. For MAP, the number and ranking of matches within a result set factor into the overall score for a submission. F-score does not factor rankings into the overall score, but considers all matches above a chosen score threshold equally. MITRE chose to use MAP as the main competition metric, as many real-world use cases for person name matching require a ranked list of results, indicating that ranking should be accounted for by the metric.

The receiver operating characteristic (ROC) metric, which plots the true positive rate against the false positive rate, also fulfills the two requirements mentioned above. The associated area under the curve (AUC) metric calculates the probability that a system will rank a

randomly chosen true positive over a randomly chosen true negative (Fawcett, 2006). MITRE chose MAP as the determining metric, rather than the analogous AUC, as MAP is commonly used in the Information Retrieval field. Additionally, in keeping with the characteristics of a good evaluation as laid out by the EAGLES working group, an evaluation should be understandable to its consumers (EAGLES, 1996). Many consumers of previous MITRE name matching evaluations understand how precision, recall, and MAP apply to their particular use cases, so we decided in favor of this metric for the Challenge.

### 2.4 Application

The MITRE Challenge application consisted of several screens, including an external login screen with a public leaderboard and several screens accessible to logged-in teams, on which they could upload results and see more detailed analysis of their result sets. The screenshot below shows one such team-accessible screen.

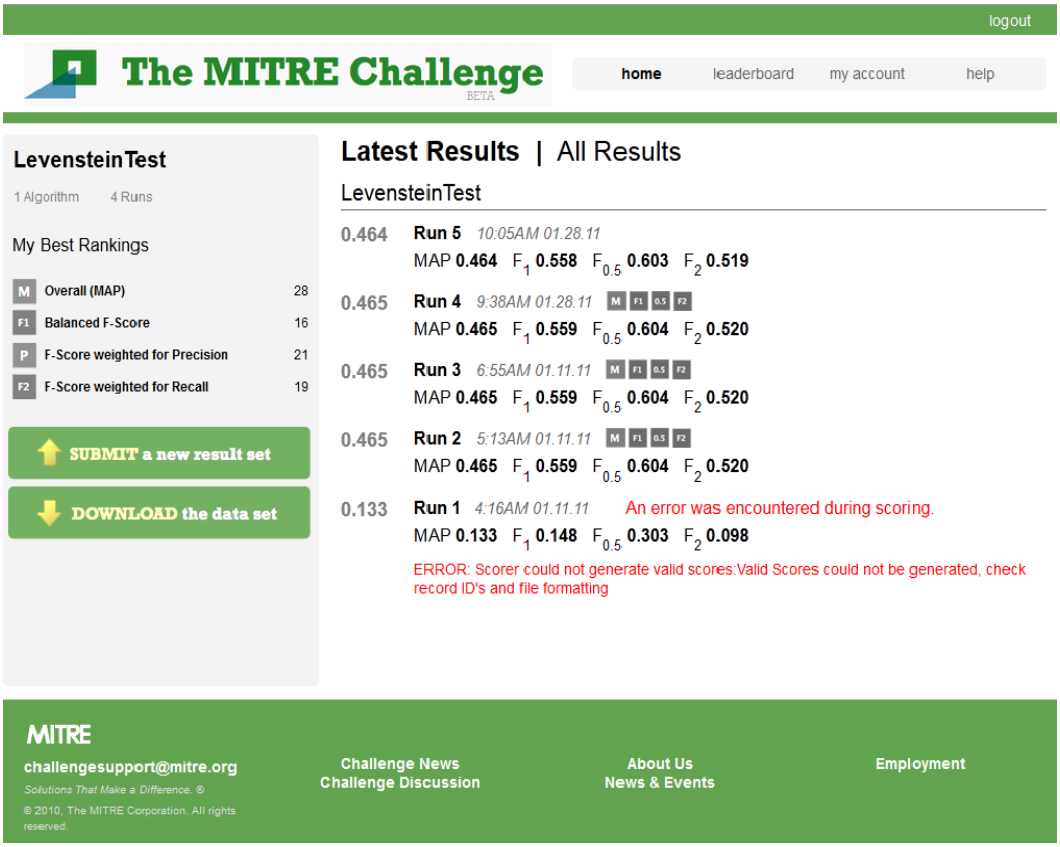


Figure 1. Participant Home Page

## 3. Data Development

As mentioned above, person names from various sources were used to create the Challenge competition data set. Those sources were chosen to mitigate privacy concerns while still providing a diverse and interesting set of person names. In addition to creating an interesting set that incorporates some of the types of person names found in real-world data sets, it was necessary to create a large data set to ensure that no team "solved" the challenge by manually comparing all possible matches.

To create the ground truth for the competition data set, MITRE employed the pooling method used in the Text REtrieval Conference (Vorhees and Harman, 2001; Vorhees 2000). In the pooling method, the data set is run through name matching software configured with a very permissive minimum matching threshold. The results of those runs are collected into an adjudication pool, then manually judged to determine whether each match is true or false. This manual adjudication process involves several human annotators, and is described in more detail in (Miller et al, 2008).

## 4. Competition Phases

The MITRE Challenge consisted of three phases of competition: the beta phase, the production phase, and the validation phase. The beta phase was similar to the beta release period in software development in that much internal testing had been done on the competition application and data set, but neither were their final state. During the beta phase, participants were encouraged to report any errors encountered. All result sets submitted during the beta phase were valid for both the beta and

production phases of the competition.

## 5. Participant Interactions

Almost 140 teams registered to participate in The MITRE Challenge. Registered teams included groups from universities, private corporations, and the US government, as well as individual participants with no particular affiliation. Teams registered from a total of twenty-one countries, including China, India, Lebanon, the US, and Netherlands. Of all registered teams, forty, representing nine countries, actively participated in the challenge by submitting one or more valid result sets.

Approximately 1/3 of the participants that submitted a result set did so within 6 hours of their initial registration. Another 1/3 did so within 3 days, and a final 1/3 took 4 days or more, with the longest time between registration and first submission being 52 days.

Most teams that submitted a valid result set submitted more than one valid result set, the most prolific of which submitted over 1,000 runs from 107 named algorithms. Although this was among the greatest number of algorithms for any team, it was not uncommon for a team to present results from multiple algorithms. In fact, 70% of the teams did submit results from more than one algorithm – on average, teams submitted results from 12 algorithms. A team from one commercial company indicated they used one team login to submit results from multiple competing teams within their organization.
Note that in this context, the decision to designate a submission as having been generated by a distinct algorithm is left to the team.

One unexpected behavior at the outset of the Challenge was that more than one team took full advantage of the Challenge guideline that allowed inclusion of 500 results per query in a result set by providing 500 results for *each* of the 8,666 queries. It seemed that they were interpreting the *limit* of 500 returns per query to mean that the best scores would be achieved by teams that returned 500 results for every query. Also in line with achieving the best scores, it was noticed that teams were exploiting the inherent features of the various contest metrics in order to maximize their scores on those metrics. This should not have been surprising, in that it is simply a real-world instance of the evaluation (and management) truism that "you get what you measure." In fact, some participants indicated during their presentations at the final technical exchange meeting that they had intentionally submitted result sets that they knew would not perform well on the main contest metric, but which they knew would perform well on one of the subsidiary metrics, given the features of those metrics.

## 6. Results

The production of phase MITRE Challenge officially concluded on 7 September 2011, and was immediately followed by the validation phase. Table 1 shows the participant team names and their highest MAP scores at the conclusion of the Challenge.

| Team Name | MAP |
|---|---|
| Mean Mr Teach | 89.6 |
| Riffraff | 89.1 |
| Beethoven | 89.0 |
| A Rose | 86.2 |
| JustForFun | 83.8 |
| Impala | 82.9 |
| 0.7 | 82.6 |
| Bach | 81.2 |
| Finite State Cola Machine | 78.8 |
| SpeedRacer | 74.3 |

Table 1. Production phase MAP scores

We performed pairwise t-tests between the highest MAP scores for each team from the production phase of the competition to determine whether any differences in scores between any two teams was significant. From this we built an NxN distance matrix, where N is the number of teams. The value of any given cell *ij* was either 0, meaning the difference in scores between teams *i* and *j* was not significant, or 1, meaning the difference was significant. From this we were able to visualize hierarchical clusters of team scores, as seen in Figure 2. In the figure, teams are ordered according to their mean score.

In the first level, the column on the far left, every team is its own cluster. In the second level, working from left to right, teams are grouped together if and only if all members of the group have no significant difference in their vector scores according to their pairwise t-test. For example, teams 1, 2, and 3 are grouped together because the difference in score was not significant between teams 1 and 2, teams 2 and 3, AND between teams 1 and 3. It is important to note that a team can be a member of two different clusters. Team 5 is grouped in a cluster with team 4 as well as in a cluster with teams 6, 7, and 8. While there was no pairwise significant difference between the scores of teams 5,6,7 and 8 and also no significant difference between the scores of teams 4 and 5, there WAS a significant difference between the scores of team 4 and those of teams 6, 7, and 8. In the third and fourth levels, the two columns on the right, clusters which shared members in the previous level are grouped together.

Each level offers a decreasing level of granularity in differentiating the meaningful difference in team scores. In the first level, each team is its own cluster and they are ranked according to mean score, where team 1 is in 1st place. In the last level, the teams have been clustered into three groups where we treat all members of group one as tied for 1st place. The t-test revealed several clusters among the competing teams: notably, it demonstrated that the performance of the top three teams on the production

test set was not statistically significantly different. Other clusters can be observed in Figure 2 as well.

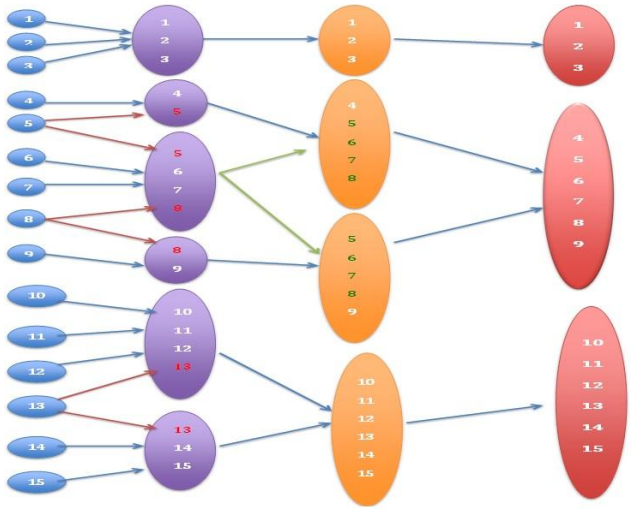

Figure 2. T-test clusters

In the validation phase of the competition teams were given a new data set. This new data followed the same format as the initial data set and had similar data characteristics, but contained entirely new names. As with the production data set, the validation set included names that contributed to the participant's score as well as names that did not. The validation query set contained 8,668 names, 168 of which contributed to the score, and 808,246 index names, 8,246 of which contributed to the score. There were 354 correctly matching name pairs in the validation data set.

Participants were required to submit up to three sets of results for the validation data set within 36 hours of receiving it. This phase of the competition served several purposes, including verification that teams could reproduce the work done to create their initial results, that this could be done in a reasonable amount of time, and that participants weren't over-fitting a solution to the production data. Satisfying the first two conditions additionally suggested that teams were in fact using computers and algorithms, as opposed to carrying out the processing task by hand.

| Team Name | MAP |
|---|---|
| Beethoven | 94.3 |
| Riffraff | 91.9 |
| A Rose | 91.7 |
| Mean Mr Teach | 91.6 |
| Impala | 90.7 |
| 0.7 | 88.9 |
| JustForFun | 88.0 |

Table 2. Validation phase MAP scores

Table 2 shows the MAP scores from the validation phase of the Challenge for those teams that placed in the top ten during the production phase and submitted a valid set of results for the validation data. Each of the teams achieved a higher MAP score during the validation phase than production phase. The participant rankings differed between the validation and production phases, with the top performer during the production phase falling to fifth place, and the third place team from the production phase jumping to first. Given the results of the paired t-test, in which the top three teams in the production phase tied for first, it could be argued that the first place team did not in fact change in the validation phase.

Overall, the results of the validation phase indicate that all of the top teams produced a replicable solution to the name matching problem, and that none of the top teams over-fit their algorithms to the production data, in which case we would have noted a drop in MAP score.

## 7. Lessons Learned

The MITRE Challenge was based on previous experience in evaluation of commercial and research name matching systems as well as knowledge of other similar competitions. Although design and execution decisions were based on these sources of prior knowledge, the team did gather many lessons from the experience of creating and running this publicly-available, externally-facing Challenge. These lessons ranged from considerations for technical design and architecture, computer security, and robustness, through issues involving communications strategies, logistics, and legal concerns. These lessons – many of which we believe will prove valuable not only for research teams seeking to run large-scale contests of the type described in this paper, but also for teams running smaller-scale evaluations – are outlined in this section.

**Communications.** The MITRE Challenge benefitted from working with both a web (UI) designer and a communications specialist. Having these team members work together to create a consistent look and feel for the Challenge site and communications materials was effective in developing a "personality" for the Challenge. This included everything from a color scheme and communications tone for the Challenge site and publicity materials to the creation of "The MITRE Challenge Squad" persona. This persona served as the principal point of interaction between interested external parties, including participating teams, and the support team for The MITRE Challenge. All external communication regarding the Challenge came in through a service e-mail account that was monitored by all members of The MITRE Challenge Squad, any of whom could respond under the MITRE Challenge Squad persona and could also read previous interactions between the Squad and the correspondent. All of this served to create a cohesive experience for parties interested in the Challenge as well as those participating in it. Communications challenges came principally in the form

of timing – allowing ample time for development and release of communications materials, coordinating and synchronizing outreach to recruit participants from a given demographic (e.g. universities), and reserving ample resources to accomplish follow-on and wrap-up communications activities necessary after the closing of the Challenge and the technical exchange meeting.

The basic mechanism used by The MITRE Challenge Squad to communicate with participating teams was a service mailing list. This worked well in that all Challenge Squad members received copies of messages sent to this service account. However, there were times that it would have been desirable to easily send outgoing communications to different subsets of the participants (e.g. participants who had achieved a certain score, those who had / had not submitted a new result set in X days, etc.). Development of this capability would likely benefit others running similar Challenges in the future, both for facilitating proactive communication with subsets of teams and for providing the ongoing support necessary during the Challenge.

**System design and Challenge logistics.** The MITRE Challenge application was designed to be as streamlined as possible, while still providing all of the necessary functionality and information to participating teams. This design was largely successful; however, it would have been useful to have had a greater number of utility, administration, and reporting functions built into the back end of the web application. As the Challenge was run, many of these functions had to be performed by the Challenge Squad manually. Further administrative functions would have been useful to provide automated tracking and prediction of rate of growth in order to allow for the proactive expansion of the virtual machine resources allocated to the Challenge. Although most of the core outward-facing functionality of Challenge-type projects can be achieved with a very lightweight design, it is recommended that such statistics and tracking backend capabilities be designed and implemented into these projects at the outset.

We also found that it was valuable to have a Beta phase – partially to work out any wrinkles in the process, and partially to verify the completeness of the ground truth data. With respect to the latter, it was useful to validate items in participant submissions that had been marked as false positives in order to identify those that should be added to the ground truth as true positives. The updated ground truth augmented in this manner replaced the beta ground truth as we moved into the production phase of the challenge. This was another area in which building more administrative functions into the backend of the challenge software would have provided overall time savings. Since there was no administrative function built into the system to facilitate the updating of team scores based on the new ground truth, it was necessary to bring the Challenge site down while all previously-submitted runs were rescored with the new ground truth.

**Testing.** As with any software intended for wide use, The MITRE Challenge site required extensive testing. Given the wide-ranging audience for Challenge-type competitions, and given the heavily bursty usage patterns they are likely to experience, test plans should be particularly thorough in order to account for all eventualities, and to include stress testing. Areas of special attention that might otherwise be overlooked include testing for "loopholes" that would allow teams to game the evaluation system and/or known features of the evaluation metrics, as well as for boundary cases allowed by Challenge guidelines (e.g. allowing 500 results for every query). Both of these were discussed in Section 5.

**Computer security.** Finally, the necessity to deal with Information Security is a reality in our times, and cannot be overemphasized. In particular, it is wise to assume that a competition that is widely publicized and is accessible to the largest possible (worldwide) audience will draw some undesired activity as well as the attention of the intended audience. Keeping this in mind, a best practice in this area would be to work with your organization's information security specialists to strike the optimal balance between information security best practices such as the "principle of least access" and the desire to provide an agreeable user experience with a low barrier to entry in order to attract the widest possible pool of appropriate participants to your competition.

## 8. Future Work

Now that the inaugural Challenge has been completed, the team is focusing on two principal thrusts: First, we are in the process of determining whether there is sufficient interest to augment this Challenge to take into account a more robust set of identity attributes in order to evaluate either multi-attribute identity matching or identity resolution. Second, we are considering other areas in which data-driven evaluation can be combined with automated calculation of metrics in order to run similar Challenges in other domains. As of the time of the writing of this abstract, no specific area has been identified for a second Challenge – but technologies both in the area of Human Language Technology and in other technology areas are being considered.

## 9. References


EAGLES Document EAG-EWG-PR.2. (1996).

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters.

Miller, Keith J., Mark Arehart, Catherine Ball, John Polk, Kenneth Samuel, Elizabeth Schroeder, Eva Vecchi and Chris Wolf (2008). An Infrastructure, Tools and Methodology for Evaluation of Multicultural Name Matching Systems. Language Resources and Evaluation Conf., Marrakech, Morocco.

Netflix Prize website. http://www.netflixprize.com/

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.