AIR®
AMERICAN INSTITUTES FOR RESEARCH®

PatentsView

Disambiguating Inventors, Assignees, and Locations

JULY 2021

Nicholas Monath | University of Massachusetts
Christina Jones | American Institutes for Research
Sarvo Madhavan | American Institutes for Research

MAKING RESEARCH **RELEVANT**

PatentsView

Disambiguating Inventors, Assignees, and Locations

JULY 2021

Nicholas Monath | University of Massachusetts
Christina Jones | American Institutes for Research
Sarvo Madhavan | American Institutes for Research

1400 Crystal Drive, 10th Floor
Arlington, VA 22202-3289
202.403.5000

www.air.org

Contents

Page

Figures

Page

Tables

Page

Introduction

Background

PatentsView is an award-winning visualization, data dissemination, and analysis platform that focuses on intellectual property (IP) data. PatentsView serves students, educators, researchers, policymakers, small business owners, and the public. It offers a unique and valuable open data platform that provides free data dissemination and value-added analyses to foster better knowledge of the IP system and drive new insights into invention and innovation.

PatentsView began in 2012 as a public–private partnership between the U.S. Patent and Trademark Office (USPTO), the U.S. Department of Agriculture, University of California–Berkeley, American Institutes for Research® (AIR®), and other partners. The beta version was released to the public in December 2015, and the full website was launched in January 2017. PatentsView's user community has grown over time. In 2019, the platform attracted an average of more than 77,000 application program interface (API) queries per day, 167 hits per day to the visualization and search interfaces, and a substantial number of direct dataset downloads.

Since its inception, the collaborators have developed and deployed several web-based tools and databases. These products include

- a web-based visualization platform with search, location, comparison, and network view functionality (https://patentsview.org/);
- an API with flexible query language and documentation (https://patentsview.org/apis/purpose);
- a bulk data download section with all data parsed from raw patent files (https://patentsview.org/download/data-download-tables);
- a query builder with an easy-to-use interface for researchers and other users to select fields and filters and receive a download link to the created .csv files (https://datatool.patentsview.org/query/); and
- a community site for patent data users to engage in discussion about patent data and post questions and comments (https://patentsview.org/welcome).

Objective

This report describes the disambiguation methodology used by the PatentsView team. It begins with a review of the steps in the process and explains how the disambiguation results are evaluated. The disambiguation process is a value-added service that is applied to USPTO's raw, publicly available data on granted patents from 1976 through quarter 2 of 2020 and is updated quarterly. The process provides unique identifiers for patent inventors, assignees (i.e., owners),

and locations. The disambiguation methodology is constantly being evaluated and updated to incorporate recent developments emerging from computer science, information science, and user experiences.

Methodology

Background

For each patent application, the USPTO collects information on the names and locations of inventor(s), assignee(s), and attorney(s) or agent(s). This information is accepted "as is" from the applicant or agent. Individuals, organizations, and representatives are not required to submit their information[1] using standardized identifiers, which leads to variation in the names and locations listed on patent pre-grant publications and on granted patent documents. For example, patent applications associated with the assignee International Business Machines will list the name for this one entity in numerous forms such as "IBM," "I.B.M.," "International Business Machines," and so forth. Misspellings are rampant throughout the location and assignee data, and common inventor names can make it difficult to understand which instances of "Tim Smith" refer to the same inventor and which do not. As a result, a disambiguation process that leverages additional patent information is needed to accurately associate unique inventor(s), organization(s), and location(s) with their patents.

PatentsView began by using an algorithm developed by a team of researchers at The Institute for Quantitative Social Science at Harvard University in 2011.[2] In 2015, PatentsView sponsored an inventor disambiguation workshop that solicited new, creative approaches to disambiguate inventors in the patent data. The most successful algorithm (authored by Nicholas Monath and Andrew McCallum from the University of Massachusetts Amherst) was subsequently integrated into the PatentsView data.

In 2017, PatentsView engaged a team at the University of Massachusetts to develop revised algorithms for the assignee and location disambiguation. Previously, the method used for disambiguating assignees was a simple edit distance algorithm. This approach had key weaknesses that could result in incorrectly grouped assignees, which is a persistent issue with assignee names consisting of a small number of characters concatenated with generic organization text (e.g., ACE Corporation) and assignees that had many characters but with small edit distances between organizations. An example would be "The United States of America as

[1] See USPTO MPEP Chapter 600 for more details (https://www.uspto.gov/web/offices/pac/mpep/mpep-0600.pdf).
[2] Lai, R., D'Amour, A., Yu, A., Sun, Y., & Fleming, Y. (2011). *Disambiguation and co-authorship networks of the U.S. patent inventor database (1975-2010).* Harvard Dataverse. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/5F1RRI

presented by the Secretary of the Navy" and "The United States of America as presented by the Secretary of the Army." The previous algorithm often assigned instances of the Army to the Navy.

The disambiguation process seeks to resolve two overlapping data issues:

1. multiple names for the same entity (inventor, assignee, or location); and
2. multiple different entities with the same name.

Computer scientists refer to these two issues as a clustering problem; that is, the need to identify which separate occurrences of an inventor, assignee, or location name (referred to as a *mention*) are the same person, organization, or location.

> ***Mentions** are defined as the separate occurrences of a name, location, or other text fields of interest that are observed in the raw patent data. For instance, "IBM" and "I.B.M." are two separate mentions.*

Overview

Figure 1 provides an overview of the disambiguation methodology. It starts with the raw patent data from the USPTO.[3] Every record in the raw data is a granted patent identified by a unique patent number, such as patent no. 9,000,000. Each patent record also has several mentions (i.e., text fields with one or more inventor names, assignee names, location names) associated with the patent.

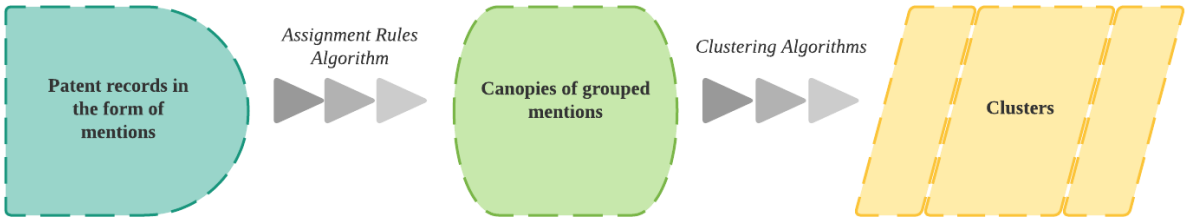
As shown in Figure 1, the first step is to group records (or, more precisely, mentions from the patent records) into "canopies." As described more in the next section, canopies are formed based on similarity rules.[4] This step addresses the tremendous volume of patent records and associated mentions in the raw data. For instance, without using canopies, evaluating the more than 15 million inventor names for potential clustering would require calculation of roughly 15 million squared pairwise comparisons. Performing hundreds of billions of similarity calculations would take too long and use too many resources to be practical.

The second step is to cluster the mentions within each canopy. This step determines which mentions represent the same inventor, same assignee, or same location within canopies. In this step, a more sophisticated similarity score is calculated that uses other cross-referencing information. For example, a "Tom Smith" who has patents in electrical engineering located in Silicon Valley is not likely to be the same "Tom Smith" who has patents in wheat varietals

[3] See https://www.uspto.gov/learning-and-resources/bulk-data-products.
[4] We are not concerned about entities with very dissimilar names. It is unlikely that "Green Solutions, Inc." is the same as "Blue Ocean, Ltd.," but we do want to compare "Green Solutions, Inc." with "Green Solvents" or "Green Solutions Company."

located in Chicago, IL, despite sharing the same name.[5] Therefore, as described later in this report, the algorithm uses a variety of cross-referencing information (e.g., technology classification codes[6] listed on patent documents and location) as part of the similarity assessment between mentions.

Figure 1. Overview of Process for Disambiguation



Step 1: Creating Canopies

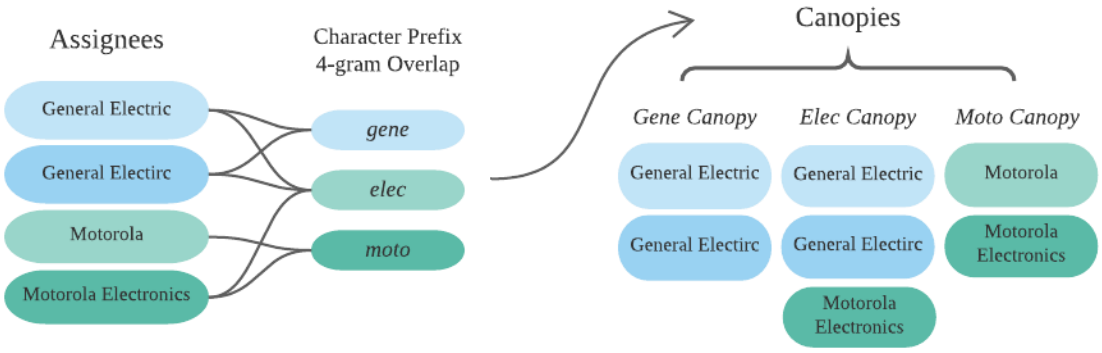
The first step in the disambiguation process is to identify records that might represent the same entity (i.e., inventor, assignee, location). This process uses a series of rules to group mentions into canopies, and each mention of an inventor, assignee, or location could be included in multiple canopies. The results are integrated across canopies after the clustering process is completed. The rules for forming canopies were developed empirically by testing many different approaches until a set of rules that divided the mentions efficiently was established.

Canopy Assignment Rules for Assignees

The algorithm uses an exact four-character overlap of the beginning of any word or name of the organization as the criteria for creating assignee canopies.[7] For example, if we have the four mentions (a) "General Electric," (b) "General Electirc," (c) "Motorola," and (d) "Motorola Electronics," the algorithm would create three separate canopies that represent three distinct sets of four characters: "gene," "elec," and "moto." Assignees with multiple word names are grouped into canopies of the first four letters of each word in the organization name (e.g. "Blue Ocean Systems" would be included in the "blue," "ocea," and "syst" canopies). Figure 2 is a visual representation of how these mentions are divided into canopies. Note that some mentions (e.g., "General Electric") appear in multiple canopies.

[5] This is a hypothetical example.
[6] Patent classifications used in the algorithm include the Cooperative Patent Classification, National Bureau of Economic Research, International Patent Classification, and United States Patent Classification.
[7] In computer science terms, an exact four-character overlap is called a character prefix 4-gram overlap.

Figure 2. Visual Representation of How Assignees Are Grouped Into Canopies



The approach described for organizations is also used for assignees who are individuals. Canopies are created based on the overlap between the first four characters of the individual name. The more in-depth approach described for inventors (in the next section) is not used for assignees who are individuals because there are relatively few mentions in this category and disambiguation using this approach is therefore not computationally taxing.

Canopy Assignment Rules for Inventors

Inventor canopies, unlike assignees, are nonoverlapping. Each inventor is assigned to a single canopy that is based on the inventor's first initial and last name. This definition of canopies is strictly more flexible than what was used in the previous PatentsView disambiguation system.

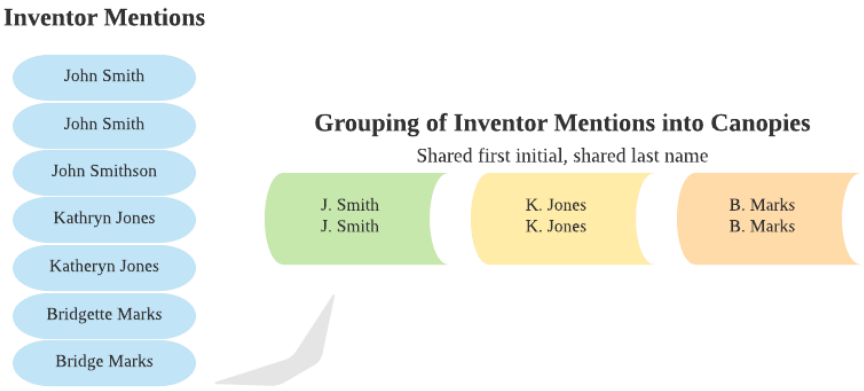
More details about our approach can be found on PatentsView.org.[8] This scheme is like the blocking techniques used by Li et al. (2014).[9] Code for our methodology can be found on GitHub.[10]

[8] Monath, N., Madhavan, S., DiPietro, C., McCallum, A., & Jones, C. (n.d.). *Disambiguating patent inventors, assignees, and their locations in PatentsView. PatentsView.* http://data.patentsview.org.s3.amazonaws.com/documents/PatentsView_Disambiguation.pdf

[9] Li, G. C., Lai, R., D'Amour, A., Doolin, D. M., Sun, Y., Torvik, V. I., ... Fleming, L. (2014). Disambiguation and co-authorship networks of the US patent inventor database (1975–2010). *Research Policy*, *43*(6), 941–955.

[10] See inventor disambiguation on GitHub (https://github.com/PatentsView/PatentsView-Disambiguation).

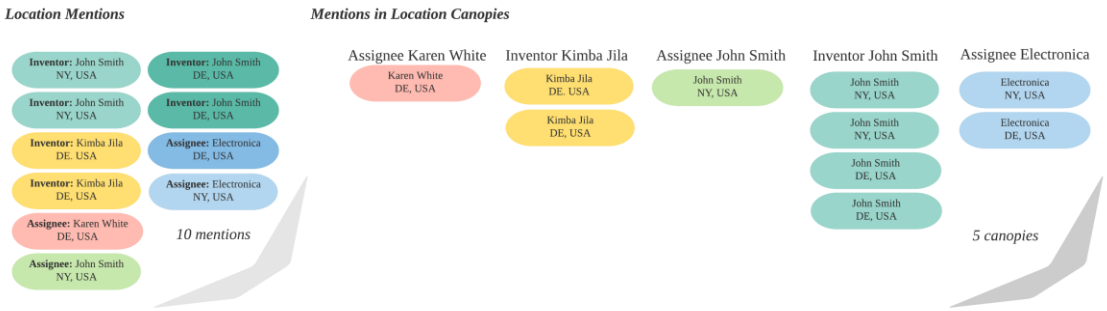
Figure 3. Process for Grouping Inventor Mentions Into Canopies



Canopy Assignment Rules for Locations

For disambiguating locations, canopies are formed after the disambiguation process for patent inventors and assignees is completed. This approach leverages the unique identifiers created in those disambiguation processes and thereby improves the information available to group location mentions. For the location canopies, we collect all the locations associated with each assignee and inventor into separate canopies. For example, with 3 million inventors and 500,000 assignees, more than 3.5 million canopies are formed. Fortunately, these canopies are relatively small. Figure 4 illustrates how location canopies are created.

Figure 4. Process for Creating Location Canopies



Step 2: Clustering

Looking within each canopy, this step groups the mentions into final clusters with unique identifiers. The unique identifiers, in the form of [patent number]-[sequence number], are the group IDs for the disambiguated raw patent records. All the mentions in these groups represent (based on the algorithm) the same entity (i.e., assignee, or inventor, or location). To cluster the mentions, a similarity score is calculated based on cross-referencing information contained in the patent records, such as patent technology classifications (e.g.,

the Cooperative Patent Classification codes) and locations. The similarity score is a rule-based metric (described next) that uses different rules for each entity type (inventor, assignee, location). With the similarity metrics calculated, a clustering algorithm determines which records are grouped together (see the section "Clustering Algorithm" below for more information).

Similarity Metrics for Assignees

We have simplified and improved our assignee similarity metric. The assignee similarity model is a pairwise model that scores the similarity between two assignees (also called pairwise similarity). Three features are used in the new assignee model: (a) name match, (b) character n-gram "bag-of-attributes" model (Table 1), and (c) PermID identifiers. These techniques are designed to identify similar text variants due to misspellings or other data limitations. For example, it is apparent that "General Electric" and "General Electirc" are likely to be the same company. The character n-gram model learns a weighted bag-of-attributes representations that characterizes two assignee names as similar if they share several sequences of characters that are unique to the two names (relative to all the other assignee names [TF-IDF weights]).

PermID[11] is a publicly available knowledge base of business entities. To measure the similarity of two assignee names, we measure to determine whether the two strings might refer to the same entity in PermID. For instance, "General Electric" and "General Electric Co." will both be "close" to the entity 4295903128 (General Electric Co.). Thus, the two will be interpreted by our PermID-based feature to be similar. However, consider "Oregon State University" and "Oregon University." These two have high textual similarity despite referring to different real-world assignees. A priori, it might be difficult for the assignee model to determine that these are different entities (especially when the word "state" appears needlessly in many other assignee names). It will be the case, however, that the two refer to different PermID identifiers and thus will be highly dissimilar by our new assignee model. Note that we use a high-precision string match to link assignee names to PermID identifiers.

Table 1. Illustration of the Bag-of-Attributes Representation

	Gree	ree	en S	Inc.		Ltd.
Green Solutions, Inc.	1	1	1	1	…	0
Green Solutions, Ltd.	1	1	1	0		1

[11] https://permid.org/

We then calculate the cosine similarity between the vector representations of the assignee mentions where the vectors contain the values for the character n-grams in the two strings. Table 2 describes the assignee similarity metrics in detail.

Table 2. Assignee Similarity Model

Feature	Description	Possible Values[a]	Feature Weight
Exact name match	Indicator for whether the names of the two assignees are exactly the same. This feature has an infinite weight for our clustering algorithm.	0 or infinity	1.0
Acronym match	Indicator for whether one assignee name is an acronym for the other, based on a dictionary of company name acronyms.	0 or infinity	1.0
"Relaxed" name match	Indicator for whether, after both names are converted to lowercase and have punctuation, spaces, and particular irrelevant words (e.g., org, ltd, co)[b] removed, the two names are the same.	0 or infinity	1.0
Prefix/suffix match	Indicator for whether the first (or last) four characters of each word in the two strings match.	0 or infinity	1.0
Name similarity	TF-IDF weighted character n-gram similarity model	Any value between 0 and 1	1.5
PermID mismatch	A binary indicator for whether the two assignee names refer to two different PermID entities.	0 or 1	-100.0

[a] The "Possible Values" column includes all the possible values that can be assigned for each feature.

[b] This list of words to be excluded (called "stop words" in the machine learning literature) was determined empirically based on review of the data.

[c] The threshold of 0.6 was determined experimentally.

[d] The thresholds of 0.89 for name similarity and 0.95 for location similarity were determined experimentally.

[e] This special condition exists because otherwise government organizations with long and mostly similar names were clustered together even though they should not be.

Similarity Metrics for Inventors

The inventor disambiguation method uses a learned linear model that determines the similarity between two sets of records. Each feature is computed as a linear function of its value and a bias term. The resulting scores from each of the features are summed to produce a final score.

For the computation of name similarity, we use a rule-based *name_match_score* function. The function is designed to determine the likelihood that names from a group of first or middle names with the same last name match. The function takes as input a list of first or middle names and a last name that is common to all the names in the group. Next steps are as follows:

1. We check the number of penalty cases. If the list of names is empty, we return a "no name penalty." Similarly, if the list of names is larger than a set maximum size, we return a "too many names penalty." Last, we check if the last name matches the common names for this group. If this is not the case, we return a "mismatch on common last name penalty."
2. After these penalty checks have been completed, we begin to compute pairwise distances between the names. For all pairs of names, we begin by checking if the first characters in the two strings match. If they do not, we increase the *firstLetterMismatches* variable by 1.
3. Next, we run our *editDistance* function to compute the difference between the given pair of names and store this in a *nameMismatches* variable.
4. The *firstLetterMismatches* value is then multiplied by an *intial_mismatch_weight* and subtracted from the score.
5. The *nameMismatches* value is multiplied by a *name_mismatch_weight,* and this value is also subtracted from the score.
6. The final score is returned by the function.

For the remaining types of features, such as coinventors, patent classifications, and lawyers, we measure the cosine similarity, Shannon entropy, and a size/quantity term (Table 3). See the Methods document[12] for more details.

Table 3. Inventor Similarity Model

Feature	Weight Name	Value	Description
First name	name_mismatch_weight	6.0	
Middle name	name_mismatch_weight	3.0	
Patent Title Embedding	cosine_similarity_weight	10.0	cosine_similarity_weight * cos_sim(fv1,fv2)
Coinventor	cosine_similarity_weight	9.5	
Coinventor	entropy_weight	0.125	
Coinventor	complexity_weight	0.5	
Assignee	cosine_similarity_weight	9.5	

Similarity Metrics for Locations

To determine similarity for locations, like for assignees, we focus on name similarity. This process assists with the handling of frequent variants and misspellings of city names. Many measures only compare records with matching states and countries. In addition to considering

[12] Monath, N., & McCallum, A. (n.d.). *Discriminative hierarchical coreference for inventor disambiguation*. University of Massachusetts Amherst Information Extraction and Synthesis Laboratory. https://s3.amazonaws.com/data.patentsview.org/documents/UMassInventorDisambiguation.pdf

the location name, the similarity measure considers the assignees and inventors associated with a given location. More detail is provided in Table 4.

Table 4. Location Similarity Features

Feature	Description	Possible Values	Feature Weight
Exact name match	Exact match across city, state, and country for both the location mentions being compared.	0 or infinity	1.0
Nonexistent location match	Indicator for whether two locations (in the same canopy, who therefore are associated with the same assignee or inventor) have the same city name and one does not exist in the MaxMind database.[a]	0 or infinity	1.0
Relaxed name match	Indicator for whether, after both locations are converted to lowercase and have punctuation and spaces removed, the two names are the same.	0 or infinity	1.0
City name similarity	Jaro-Winkler[b] similarity between city names, if the state and country are the same. 0 if the state and country are not the same. The Jaro-Winkler similarity is a common measure of how similar two strings are.	Any value between 0 and 1	1.0
Name incompatibility	A binary indicator for whether the Jaccard similarity of the assignee names is less than 0.75.[c]	0 or 1	–10.0
Overwhelming number of records match	When comparing similarity between two location clusters during the clustering process, if the two locations' city names are the same (using a relaxed match as described above) and one cluster has more than 1.5 times as many records associated with it as the other.[d]	0 or infinity	1.0

[a] This measure seeks to capture specifically records in which the wrong state is written for a location but the same inventor/assignee has patented in that city and entered the correct state on a different patent.

[b] Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods* (pp. 354–359). American Statistical Association. https://files.eric.ed.gov/fulltext/ED325505.pdf

[c] The threshold of 0.6 was determined experimentally.

[d] This metric seeks to merge the oddly formatted city names (with strange capitalization or punctuation) with the more standard formatted, and thus more common, name of that city.

Clustering Algorithm

The next step in the disambiguation process is to use a clustering algorithm to group mentions together. We use hierarchical agglomerative clustering to cluster inventors, assignees, and locations.

1. Compute the similarity score for each pair of mentions within each canopy.

2. Group together the two most similar records. For assignees, if there are more than 1,000 records in the canopy, only a sample are compared, as described above for inventors, to reduce computational overhead.
3. Repeat the comparisons between all mentions and the newly formed clusters from step 2.
    a. For assignees, the similarity between the cluster and any other mention is defined as the maximum similarity between any element in the cluster and the mention with which it is being compared.[13]
    b. For locations, the similarity between the cluster and any other mention is defined as the similarity between the concatenation of all mentions in the cluster and the mention with which it is being compared. Groups are represented by the canonical name when comparing similarity measures between a group and a mention.
4. Each time records are clustered together, they form a node in the group that we are creating.
5. After all the records in a canopy are formed into a group, the final clusters are any groups whose similarity score exceeds the empirically determined threshold.

All records in a canopy become clustered based on similarity scores. Each time records are clustered together, they form a node in a tree. After all records are formed into a tree, the final clusters are a subtree whose similarity score exceeds a determined threshold. To incrementally add data to a clustering produced by hierarchical agglomerative clustering, we use agglomerative clustering's incremental variant, Grinch.[14]

Evaluation

Evaluation is an important part of the process of creating and understanding disambiguation algorithms. It allows users to assess the quality of the output and it allows the PatentsView team to gauge the value of proposed improvements to the algorithm.

However, this endeavor is challenging because no "gold standard" or "ground truth" dataset has all the properties one would like to properly evaluate the accuracy of the clusters (i.e., disambiguated groups of *mentions*). As an alternative, we used nine different manually labeled evaluation datasets with samples from the full set of patent data.[15]

[13] This is called single linkage.
[14] Monath, N., Kobren, A., Krishnamurthy, A., Glass, M. R., & McCallum, A. (2019). Scalable hierarchical clustering with tree grafting. In *KDD '19: The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 1438–1448). Association for Computing Machinery. http://bit.ly/grinch_paper
[15] Because of the high number of records, it is not possible to code all the patent data manually.

We used two approaches for evaluating the results of the inventor, assignee, and location disambiguation algorithms. The first approach uses standard metrics found in the computer science literature (precision, recall, and F1) to determine the number of mentions correctly and incorrectly classified. The second approach uses a metric derived by the PatentsView team that focuses on the number of differently spelled names that were correctly and incorrectly assigned. This evaluation dataset was constructed to address cases that were difficult to disambiguate. All evaluation datasets described in this evaluation are publicly posted.[16]

Evaluation Datasets

The following describes the nine datasets used for the disambiguation evaluation[17]:

1. **NBER.** The National Bureau of Economic Research provides disambiguated assignee data.[18] These data are created semiautomatically with manual correction and labeling of assignee coreference decisions produced by string similarity. We grouped the assignee mentions by four-letter prefixes and focused on five prefix groups {Moto, Amer, Gene, Solu, Airc} that were both common and ambiguous.[19]
2. **PatentsView assignee.** The PatentsView team created a hand-labeled set of disambiguated assignee records. The data were created by sampling records of each assignee type (universities, federal government entities, private companies, states, and local government agencies). We used those records as queries for annotators to find all other records referring to the same assignee. Team members annotated the labeled records according to string similarity. In cases where an identity could not be confirmed or was uncertain, annotators did not create a link. We intended this dataset to have a larger coverage of name varieties of the entities than the NBER dataset, which was important for us to evaluate the more difficult-to-disambiguate cases. Annotators attempted to label parent companies separately from subsidiaries, but the process was more likely to associate similarly named child and parent companies than more distinctive ones.
3. **Engineers and scientists.**[20] This sample from the preexisting, public Png LinkedIn Patent Inventor FIVES dataset (produced by Dartmouth) associated more than 14,000 inventors with all of their patents.

[16] For data and results, see https://s3.amazonaws.com/disambiguation-eval/pv-eval-data.zip.
[17] All evaluation datasets and evaluation code can be downloaded from PatentsView at https://s3.amazonaws.com/disambiguation-eval/pv-eval-data.zip.
[18] For data from NBER, see https://sites.google.com/site/patentdataproject/Home/downloads.
[19] By ambiguous, we mean that several assignees share that prefix. For instance, if we were to examine the prefix "ibm," this would not give us much information to evaluate because this prefix only has one assignee.
[20] FIVESProject: http://five.dartmouth.edu/datasets

4. **Israeli inventor dataset.** This sample from a hand-labeled dataset tracks the patenting activity of 6,000 Israeli inventors who patent in the United States.[21]
5. **Academic life science dataset.** This dataset of academics in the life sciences field was created by the PatentsView team.
6. **Particularly challenging academic life science dataset.** A hand-created set of particularly difficult to disambiguate names from the Academic Life Science Dataset.
7. **PatentsView inventor.** The PatentsView team created a hand-labeled set of inventor records to capture how the system performs on very common last names (names that are more likely to have a greater number of overlapping names). We selected three common surnames: "Moore," "Peterson," and "Chen" and one less common surname "Maak" and labeled 93 records that corresponded to 67 unique entities using assignee, location, year, lawyer, and other inventor information in the PatentsView data.
8. **PatentsView inventor location.** The PatentsView team hand-labeled the disambiguation of inventor locations. We focused on labeling city names that appear in multiple countries and states. We then selected inventors who have these location records as a basis for annotation. Note that annotation of inventor records is considerably more challenging than assignees because many unique inventors share the same name. To create the dataset, annotators needed to consider patent topic, assignee, and other patent information.
9. **PatentsView assignee location.** The PatentsView team hand-labeled the disambiguation of assignee locations. As in the PatentsView assignee, we focused on labeling city names that appear in multiple countries and states. We selected assignees that have location records as a basis for annotation. Compared with inventor location data, assignee data can be more easily verified, and groups of records could be annotated together quickly.

Table 5 describes the evaluation datasets by number of records and unique entities. Table 6 describes the name variation dataset by minimum, maximum, and average number of name entities for the assignee evaluation datasets.

Table 5. Descriptive Statistics for Each Evaluation Dataset

	Number of Records	Number of Unique Entities
NBER assignee	238,398	7,236
PatentsView assignee	371,599	111
Engineers and scientists	10,595	1,821

[21] Trajtenberg, M., & Shiff, G. (2008). *Identification and mobility of Israeli patenting inventors*. Discussion Paper No. 5-2008. The Pinhas Sapir Center for Development, Tel Aviv University.

	Number of Records	Number of Unique Entities
Israeli inventor dataset	2,060	742
Academic life sciences dataset	41,347	4,743
Particularly challenging academic life sciences dataset	9,396	1,150
PatentsView inventor	93	67
PatentsView location (inventor)	368	49
PatentsView location (assignee)	678,890	111

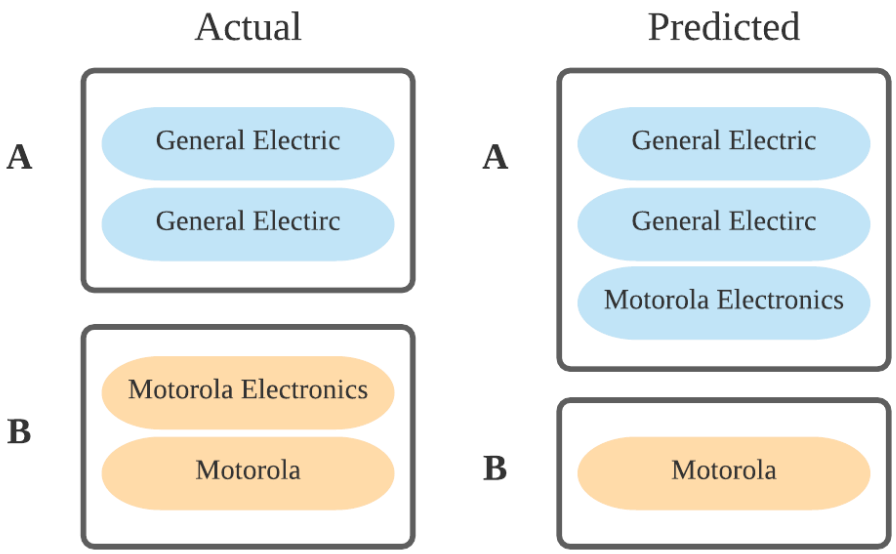
Table 6. Descriptive Statistics for the Assignee Datasets for the Name Variation Metric

	Minimum Number of Entity Name Variations	Maximum Number of Entity Name Variations	Average Number of Entity Name Variations
NBER assignee	1	166	2.09
PatentsView assignee	1	305	35.92

Evaluation Metrics

To evaluate the quality of the disambiguation algorithms, we created metrics to compare the predicted clusters produced by the algorithm against the actual clusters from our labeled evaluation datasets. An example is shown in Figure 5. Some records were correctly grouped together. For example, “General Electric” and “General Electirc” both were predicted to be in cluster A together, but the other records were not. “Motorola Electronics” is in cluster A instead of cluster B.

Figure 5. Example of Partially Incorrect Clustering



To quantify predictions, for each predicted cluster we calculate the number of records that are correctly in the cluster (true positives), the number of records incorrectly included in the cluster (false positives), and the number of records incorrectly left out of the cluster (false negatives).

In Figure 5, cluster A has two true positives ("General Electric" and "General Electirc" are both correctly in cluster A), one false positive ("Motorola Electronics" is in cluster A and should not be), and no false negatives (all records that should be in cluster A are in cluster A). For cluster B, we have one true positive ("Motorola" is correctly in cluster B), no false positives (there are no records in the cluster that should not be), and one false negative ("Motorola Electronics" should be in cluster B and is not).

Three metrics are used to score the balance among true positives, false positives, and false negatives: precision, recall, and pairwise F1 scores. These metrics, explained next, emphasize different aspects of correctness of the clusters that the algorithm creates.

***Precision** of the clustering algorithm measures what fraction of the mentions predicted to be in a cluster are supposed to be there (i.e., true positives).*

Mathematically, this is:

$$\frac{\#\text{ true positives}}{\#\text{ true positives} + \#\text{ false positives}}$$

These metrics measure, intuitively, how precise the algorithm is; that is, how many records that we think belong together actually belong together.

***Recall** of the clustering algorithm measures what fraction of the records should have been in the cluster and were predicted to be in that cluster.*

Mathematically, this is:

$$\frac{\#\text{ true positives}}{\#\text{ true positives} + \#\text{ false negatives}}$$

F1 Score. Recall has a natural tradeoff against precision. If we want to be very careful to only put a record into a cluster if we are sure it belongs (i.e., to have a high precision), then we end up not putting some records into clusters where they actually belong (i.e., we will lower our recall). If we want to be really sure not to leave any record out of a cluster that it should be in (i.e., to have a high recall), then we will put some records into clusters where they do not actually belong (i.e., we will lower our precision). To balance precision against recall, it is common to use the F1 metric.

> ***F1** score of a clustering algorithm is the harmonic mean of precision and recall.*

Mathematically, this is:

$$2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

Evaluation Methodology

We took two major approaches for evaluating the disambiguation algorithms. The first approach evaluated the algorithm based on the standard precision, recall, and F1 metrics; essentially, we examined the number of mentions correctly and incorrectly classified (Table 7). This assessment shows how well the algorithms are performing (although the test cases were purposely selected to be more difficult to disambiguate than the average PatentsView mention).

Table 7. Evaluation Results That Compare the Precision, Recall, and F1 Scores for the Pairwise F1 Metric on PatentsView Disambiguation Methodologies

		Evaluation Datasets			
		NBER Assignee	PatentsView Assignee	Location (Inventor)	Location (Assignee)
Disambiguation methodology for PatentsView data updated through 05/28/2018	Precision	0.838	0.977	1.000	0.998
	Recall	0.971	0.977	1.000	0.975
	F1	0.900	0.977	1.000	0.986
Disambiguation methodology for PatentsView data updated through 11/27/2018	Precision	0.957	1.000	1.000	0.999
	Recall	0.928	0.981	0.980	0.990
	F1	0.942	0.991	0.990	0.994

		Evaluation Datasets			
		NBER Assignee	PatentsView Assignee	Location (Inventor)	Location (Assignee)
Disambiguation methodology for PatentsView data updated through 6/14/2021	Precision	0.992	1.000	1.000	1.000
	Recall	0.991	0.9795	0.951	0.994
	F1	0.992	0.9896	0.975	0.997

For assignees, we additionally compare the pairwise F1 metric normalized by the number of occurrences of each unique assignee name. That is, we measure the F1 statistic where the atomic unit is a unique string rather than a record in the database. We report these results in Table 8.

Table 8. Evaluation Results That Compare the Precision, Recall, and F1 Scores for the Pairwise F1 Metric Normalized by Unique Assignee Names

		PatentsView Assignee
Disambiguation methodology for PatentsView data updated through 11/27/2018	Precision	0.998
	Recall	0.331
	F1	0.497
Disambiguation methodology for PatentsView data updated through 6/14/2021	Precision	0.992
	Recall	0.372
	F1	0.541

The second approach calculated the same metrics but focused on the number of distinct entity names instead of mentions. The core idea is that our algorithm could perform well using the overall metrics without being good at disambiguating. For example, say we have 300 mentions of "Green Farms Corn Technologies," one mention of "Green**s** Farms Corn Technologies," and two mentions of "Green Farms Corn Technology." If our algorithm correctly clustered the 300 "Green Farms Corn Technologies" together but excluded the other three mentions, it would have a very high recall (300/300+3 = 0.99). However, the algorithm would not have succeeded in clustering any of the entities with differently spelled names.

To measure this aspect of disambiguation performance, we examined specifically what we call Name Variation evaluation metrics. Of the set of different spellings of a name, this metric measures how many of the name variants were included in the predicted clusters. In the example in the previous paragraph, our recall would now be much lower because we correctly identified one name (one true positive) and falsely excluded two names (two false negatives) for a recall of 1/3 = 0.3333. The results for this more challenging metric are shown in Table 9.

Table 9. Evaluation Results Comparing the Average, True Positives, False Positives, False Negatives, and F1 Scores for the Name Variation Evaluation Metric on PatentsView Disambiguation Methodologies

		Evaluation Datasets			
		NBER Assignee	**PatentsView Assignee**	**Location (Inventor)**	**Location (Assignee)**
Disambiguation methodology for PatentsView data updated through 05/28/2018	Average (lower better)	2.49 ± 15.25	41.73 ± 84.89	0.65 ± 0.56	1.94 ± 2.70
	True positive	7,305	2,036	26	229
	False positive	10,155	2,682	0	18
	False negative	7,841	1,951	32	197
	F1 (higher better)	0.448	0.468	0.619	0.681
Disambiguation methodology for PatentsView data updated through 11/27/2018	Average (lower better)	1.90 ± 7.02	31.14 ± 54.29	0.69 ± 0.54	0.78 ± 1.30
	True positive	6,716	2,207	25	352
	False positive	5,312	1,676	1	13
	False negative	8,430	1,780	33	74
	F1 (higher better)	0.494	0.561	0.595	0.890
Disambiguation methodology for PatentsView data updated through 06/14/2021	Average (lower better)	2.9527± 10.0522	34.369± 49.005	0.551 ± 0.536	1.117 ± 4.050
	True positive	7,698.0	2,206.0	32	352
	False positive	13,915.0	2,051.0	1	13
	False negative	7451.0	1,764.0	26	74
	F1 (higher better)	0.419	0.536	0.703	0.890

Next Steps

This report examined the current disambiguation methodology used by the PatentsView team. The algorithms for PatentsView disambiguation are continuously being evaluated by the PatentsView team and PatentsView users.

To inform future upgrades to the algorithms, the USPTO and AIR hosted the 2021 Symposium on Entity Resolution to bring together computer scientists, information scientists, economists, and other interested researchers, policymakers, and thought leaders to discuss state-of-the-art approaches to and current applications of entity resolution, particularly focused on applications to patents. For more information, see https://patentsview.org/entityres.

AIR®

AMERICAN INSTITUTES FOR RESEARCH®

Established in 1946, the American Institutes for Research® (AIR®) is a nonpartisan, not-for-profit organization that conducts behavioral and social science research and delivers technical assistance both domestically and internationally in the areas of education, health, and the workforce. AIR's work is driven by its mission to generate and use rigorous evidence that contributes to a better, more equitable world. With headquarters in Arlington, Virginia, AIR has offices across the U.S. and abroad. For more information, visit www.air.org.

MAKING RESEARCH **RELEVANT**

AMERICAN INSTITUTES FOR RESEARCH

1400 Crystal Drive, 10th Floor

Arlington, VA 22202-3289 | 202.403.5000

www.air.org

LOCATIONS

Domestic: Arlington, VA (HQ) | Sacramento and San Mateo, CA | Chicago, IL | Indianapolis, IN | Waltham, MA | Rockville, MD | Chapel Hill, NC | Austin, TX

International: Ethiopia | Haiti

15521_07/21