Research Policy 44 (2015) 1672–1701

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Seeing the non-stars: (Some) sources of bias in past disambiguation approaches and a new public tool leveraging labeled records

Samuel L. Ventura [a], Rebecca Nugent [a], Erica R.H. Fuchs [b,*]

[a] Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, United States
[b] Department of Engineering and Public Policy, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, United States

ARTICLE INFO

Article history:
Received 21 July 2014
Received in revised form 2 December 2014
Accepted 28 December 2014
Available online 7 February 2015

Keywords:
Record linkage
Disambiguation
Patents
Supervised learning
Random forests

ABSTRACT

To date, methods used to disambiguate inventors in the United States Patent and Trademark Office (USPTO) database have been rule- and threshold-based (requiring and leveraging expert knowledge) or semi-supervised algorithms trained on statistically generated artificial labels. Using a large, hand-disambiguated set of 98,762 labeled USPTO inventor records from the field of optoelectronics consisting of four sub-samples of inventors with varying characteristics (Akinsanmi et al., 2014) and a second large, hand-disambiguated set of 53,378 labeled inventor records corresponding to a subset of academics in the life sciences (Azoulay et al., 2012), we provide the first supervised learning approach for USPTO inventor disambiguation. Using these two sets of inventor records, we also provide extensive evaluations of both our algorithm and three examples of prior approaches to USPTO disambiguation arguably representative of the range of approaches used to-date. We show that the three past disambiguation algorithms we evaluate demonstrate biases depending on the feature distribution of the target disambiguation population. Both the rule- and threshold-based methods and the semi-supervised approach perform poorly (10–22% false negative error rates) on a random sample of optoelectronics inventors – arguably the closest of our sub-samples to what might be expected of the majority of inventors in the USPTO (based on disambiguation-relevant metrics). The supervised learning approach, using random forests and trained on our labeled optoelectronics dataset, consistently maintains error rates below 3% across all of our available samples. We make public both our labeled optoelectronics inventor records and our code to build supervised learning models and disambiguate inventors (see http://www.cmu.edu/epp/disambiguation). Our code also allows users to implement supervised learning approaches with their own representative labeled training data.

1. Introduction

Disambiguation, or the process of linking records of unique individuals or entities within a single data source, is a subset of the broader "Record Linkage" field, which is generally used to link records of unique individuals or entities across multiple data sources. In 1969, Ivan Fellegi and Alan Sunter introduced the first mathematical model for record linkage (Fellegi and Sunter, 1969); this model is still the basis for many of the most common approaches to record linkage used today. In the field of technology, innovation, and entrepreneurship (TIE), record linkage and disambiguation are used to link records of assignees (the companies, organizations, individuals, or government agencies to which a patent is assigned) and, notably, to link records of inventors in the United States Patent and Trademark Office (USPTO) database. However, many USPTO disambiguation approaches fail to take advantage of the latest methodological advancements in statistics, such as adaptations of the Fellegi and Sunter (1969) approach for record linkage (e.g. Fleming et al., 2007; Lai et al., 2009). More importantly, many existing USPTO inventor disambiguation algorithms often use ad hoc weights, thresholds, and decision rules to determine which records should be linked (e.g. Lai et al., 2009) instead of leveraging information from "labeled inventor records," or USPTO inventor records for which the true identity of the inventor is known, during disambiguation. Such approaches may introduce prevalent and systematic errors in the disambiguation results, which might be avoided by leveraging information from labeled inventor records.

Using two sets of labeled USPTO inventor records from different scientific and institutional contexts (98,762 records from the field of optoelectronics consisting of four sub-samples of inventors

* Corresponding author. Tel.: +1 412 268 1877.
E-mail addresses: sventura@stat.cmu.edu (S.L. Ventura), rnugent@stat.cmu.edu (R. Nugent), erhf@andrew.cmu.edu (E.R.H. Fuchs).

http://dx.doi.org/10.1016/j.respol.2014.12.010

with varying characteristics (Akinsanmi et al., 2014) and 53,378 records corresponding to superstar academics in the life sciences with patents (Azoulay et al., 2012)), we make two contributions to the TIE field and the USPTO inventor disambiguation literature. First, we evaluate three commonly used inventor disambiguation approaches (Fleming et al., 2007; Lai et al., 2009, 2014), arguably representative of the range of approaches used to disambiguate USPTO inventors to-date, to determine the rates of false positive and false negative errors in their disambiguation results. These three approaches include two examples of unsupervised, rule- and threshold-based approaches: The first, Fleming et al. (2007), is similar also to past approaches such as those by Singh (2005) and Jones (2005). The second, Lai et al. (2009), is similar also to past approaches such as those by Trajtenberg et al. (2006), Lissoni et al. (2006), and Miguelez and Gomez-Miguelez (2011). We also evaluate one semi-supervised learning algorithm trained on statistically generated artificial labels, Lai et al. (2014). Second, we contribute the first supervised learning approach to the USPTO inventor disambiguation problem. Here, we build and evaluate statistical classification models for inventor disambiguation using information from the labeled inventor records to inform the algorithm. We then compare the disambiguation results of the best-performing classification model to the unsupervised and semi-supervised approaches described above. For the purposes of this study, we consider false negative errors and false positive errors to be equally unfavorable in the results of any disambiguation algorithm, though there are some contexts where one type of error may be favorable to the other (Fegley and Torvik, 2013). We define a splitting metric to assess false negatives where a single inventor is "split" into multiple inventor IDs, and a lumping metric to assess false positives where multiple inventors are "lumped" into one inventor ID. Our goal is to consistently achieve a balance of both low splitting errors and low lumping errors across the range of labeled sub-samples with different disambiguation features available to us. Here, consistent performance across contexts is equally important to balance, as a disambiguation algorithm that performs inconsistently across contexts would provide results that suggest differences across, for example, institutional or industrial contexts (or particular types of inventors) that are created by the disambiguation algorithm rather than being a reality in the original data. To summarize, we choose to pursue *consistency* across contexts and *balanced* splitting and lumping in the interest of pursuing the most generally useful disambiguation results across the wide range of research questions and contexts that might be explored using the data, rather than optimizing the results to what might be most useful to a particular context or question.

While the three past disambiguation algorithms we evaluate perform well in certain contexts, they perform inconsistently (e.g. demonstrate biases) across contexts depending on the feature distribution of the target disambiguation population. We find that the Fleming et al. (2007) has high splitting rates when evaluated against both the optoelectronics (OE) and the academic life sciences (ALS) labeled datasets. Lai et al. (2009) (based on publicly posted results where the algorithm is run on the full USPTO) relatively accurately disambiguates the set of academics in the life sciences with patents, but continues to display high splitting rates for disambiguating optoelectronic inventors. An important difference between the OE and ALS datasets is that in the ALS dataset, inventors appear to submit relatively consistent information to the USPTO (something we hypothesize may be more likely for academics and non-mobile inventors), include their middle initial, and are primarily U.S.-based. In contrast, in the optoelectronics dataset, middle names and other fields are frequently missing, and the proportion of U.S. inventors is (as in the full USPTO) only approximately half of all inventors in the sample, making it more difficult to disambiguate. The semi-supervised Lai et al. (2014) algorithm (again, based on publicly posted results where the algorithm is run on the full USPTO) at first appears to outperform all other inventor disambiguation algorithms, including slightly outperforming our supervised learning approach, when evaluated on the full optoelectronics and the full academic life sciences datasets. However, when we unpack the performance of the rule- and threshold-based methods and the semi-supervised Lai et al. (2014) algorithm on individual subsets of the OE dataset we once again find that they performs inconsistently across contexts: Specifically, they perform particularly poorly on a critical subset of the optoelectronic database – our random sample of optoelectronics inventors – which is arguably the closest of our sub-samples to what might be expected of the majority of inventors in the USPTO (based on measurable disambiguation-relevant metrics). Here, these algorithms yield splitting rates from 10% (the Lai et al. (2014) semi-supervised approach) to over 20% (the (Fleming et al., 2007) rule- and threshold-based approach). In contrast, the supervised learning approach, using random forests (Breiman, 2001) trained on the OE dataset, consistently maintains error rates below 3% across all of our available samples, including the random sample of optoelectronic inventors.

Our results suggest it important for the TIE field to continue to pursue disambiguation approaches that are consistent across disambiguation contexts with varying features. We also show that to assess past theoretical work using the disambiguated results from the algorithms evaluated in this paper or other algorithms with similar approaches, it will be important to look at the suitability of the research contexts and questions to the chosen disambiguation approach's respective strengths and weaknesses. The performance of our algorithm on additional USPTO datasets (whether other industrial and institutional contexts or the full USPTO database) is inevitably limited by the features of the labeled USPTO inventor records to which we had access. Incorporating labeled records with useful features (including detailed information on non-matches) from alternative samples will likely improve our random forests algorithm's ability to disambiguate additional USPTO datasets, since this will allow samples of records with different features to be represented and accounted for in our models. To continue to improve inventor disambiguation in the USPTO and interpretation of research leveraging the disambiguation results of past disambiguation algorithms, it will be important to continue to evaluate existing and future approaches on other sets of labeled inventor records, both to identify additional areas of potential bias in existing models upon which past papers have been based and to evaluate and improve future supervised and semi-supervised learning models used for USPTO inventor disambiguation. It is also imperative that the field moves towards requiring authors to publish as part of their theoretical papers the disambiguation approach used to generate the data upon which the theory is built, including a discussion of where that disambiguation approach may have biases.

We make public (http://www.cmu.edu/epp/disambiguation) all code and labeled inventor records for our disambiguation process, for use by both the USPTO research community and the broader disambiguation and record linkage communities.[1] Our code allows users the flexibility to specify their own blocking criteria to support applying our algorithm to databases of different size, build supervised learning models on their own labeled training data representative of their target population for disambiguation, and adjust the disambiguation results depending on their desired prevalence of false positive and false negative matching errors (in accordance with their particular research question). In providing public access not only to our algorithm but also to our extensive

[1] Several past authors have also released software for record linkage and disambiguation, including Goiser and Christen (2006), Elfeky et al. (2003), and Christen (2008), among others.