Author Name Disambiguation in MEDLINE Based on Journal Descriptors and Semantic Types

Dina Vishnyakova
Roche Pharmaceutical Research and Early Development, pRED Informatics, Roche Innovation Center, Basel, Switzerland and Swiss Institute of Bioinformatics, Switzerland
Dina.Vishnyakova@roche.com

Raul Rodriguez-Esteban
Roche Pharmaceutical Research and Early Development, pRED Informatics, Roche Innovation Center, Basel, Switzerland
raul.rodriguez-esteban@roche.com

Khan Ozol
Novartis Pharmaceuticals, Basel, Switzerland
khan.ozol@novartis.com

Fabio Rinaldi
Institute of Computational Linguistics University of Zurich and Swiss Institute of Bioinformatics, Switzerland
fabio.rinaldi@uzh.ch

Abstract

Author name disambiguation (AND) in publication and citation resources is a well-known problem. Often, information about email address and other details in the affiliation is missing. In cases where such information is not available, identifying the authorship of publications becomes very challenging. Consequently, there have been attempts to resolve such cases by utilizing external resources as references. However, such external resources are heterogeneous and are not always reliable regarding the correctness of information. To solve the AND task, especially when information about an author is not complete we suggest the use of new features such as journal descriptors (JD) and semantic types (ST). The evaluation of different feature models shows that their inclusion has an impact equivalent to that of other important features such as email address. Using such features we show that our system outperforms the state of the art.

1 Introduction

A frequent task for researchers is searching for relevant publications or citations. These resources are often queried by the name of an author. According to Dogan et al. (2009) queries based on Author Name are most frequent in PubMed and make approximately 36% of all queries. However, author names can be highly ambiguous, which complicates any author search and posterior analysis. Although some online literature resources partially disambiguate author names-for example, PubMed started to rank authors according to the likelihood that they are relevant to a user author name query since 2012 (Liu et al., 2014) - this is not yet an established practice. Moreover, when querying for particular topics or subjects in PubMed it is very challenging for a user to figure out the key authors relevant to the query and PubMed does not offer any aid in that respect.

Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2016)

Several articles regarding Author Name Disambiguation (AND) solutions in MEDLINE have been published, e.g. (Smalheiser and Torvik, 2009; Torvik et al., 2005; M. Song et al., 2015; Liu et al., 2014; Li et al., 2012; Warner, 2010; Treeratpituk and Giles, 2009). However, the AND problem is not yet satisfactorily solved. Alternatively, unique identifiers for authors such as those from *Scopus* or *ORCID* (Haak et al., 2012) have been created in order to disambiguate names in publications. However, a unique author identifier is not a requisite for publishing (Smalheiser and Torvik, 2009). Moreover, some existing unique identifiers assigned to authors by citation or abstract databases such as *Scopus* or *arXiv Author ID* (Warner, 2010) are based on an automatic information extraction mechanism and often are not validated by the authors themselves, and therefore can contain errors.

The majority of the methods described in (Torvik et al., 2005; M. Song et al., 2015; Liu et al., 2014; Li et al., 2012; Warner, 2010; Treeratpituk and Giles, 2009) base their disambiguation methods on author personal data from MEDLINE records such as name, affiliation, co-authorship and e-mail address (Torvik et al., 2005; M. Song et al., 2015; Liu et al., 2014; Cota et al., 2010). While information regarding an author's last name and first name is an essential part of a scientific article, information regarding the author's affiliation is not always provided by MEDLINE. As an example, (Liu et al., 2014) mentions that information about affiliation was available only in 53% of the publications they considered. Beyond personal data, information such as MeSH terms and keywords has also been used for disambiguation. According to (Liu et al., 2014), the availability of MeSH terms in MEDLINE is ~ 91%, which is larger, in the sense of publication coverage, than the availability of affiliation information.

Commonly, disambiguation methods estimate author publications within the same "equivalence set," where each set is defined by all the authors that share the same last name and first initial. This means that author publications need to be grouped first by last name and first initial (Torvik and Smalheiser, 2009; Liu et al., 2014; M. Song et al., 2015; Cota et al., 2010). We will refer to such equivalence sets as "namespaces." Thus, identifying the namespaces is the initial and most important step for AND. Thereafter, the methods to disambiguate authorship can vary depending on the features used, which are selected to calculate inter-publication similarity. Evidently, the process of assigning an author's publication to a namespace may affect the overall results of the disambiguation.

Usually authors tend to publish their work in specific journals, conferences, workshops, etc. depending on the topics of the journals and the research domain of the author. However, in the era of translational research it becomes problematic to strictly define which topics belong to which author. This can be done, for example, through the analysis of the keywords or MeSH terms used in the author's publications or by creating author-journal similarity profiles (Torvik et al., 2005; Y. Song et al., 2007). However, when the paper has several authors, the identification of the main topics of interest of each author/co-author becomes challenging. Moreover, publications written by specialists from different domains collaborating on a common project may include key terms from different fields/domains. We propose, instead, to use journal descriptors (JDs) to aid in AND instead of keywords mentioned in the publication. The JDs add more detail by describing the different specialties associated to the articles. They can identify not only the main domain of an article but also secondary ones.

2 Methods

In this section we describe the features we used and their provenance for creating "author profiles." By an author profile we mean an array with the following values associated to a particular publication: 1) Last name, 2) First name, 3) Initials, 4) Publication ID (PMID), 5) Year of publication, 6) Language of the publication, 7) Title of the publication, 8) Abstract, 9) MeSH terms, and 10) Affiliation.

2.1 MEDLINE information

Initially, all information available in MEDLINE regarding the author of each publication was extracted. This information includes the following: 1) last name, 2) full first name, 3) initials, 4)

affiliation, 5) co-authors, 6) order of the author in the author list, 7) language of the publication, 8) MeSH terms, 9) abstract and 10) title.

Information regarding organization, city, country as well as email address were extracted from the author's affiliation. To extract the email address from the affiliation, a regular expression was used. In order to extract the organization name, the Stanford named-entity recognizer (NER) based on the 7-class model (Finkel et al., 2005) was used. This model has been trained on the MUC6 (https://catalog.ldc.upenn.edu/LDC2003T13) and MUC7 training data (https://catalog.ldc.upenn.edu/LDC2001T02). The model recognizes location, organization, person, date, money, percent and time information in text. The choice of this NER algorithm can be explained by its better performance compared to OpenNLP (Dlugolinský et al., 2013). Since affiliation information is usually represented as a short text string it was important to choose the NER model which could recognize entities with a better accuracy in such strings. A preliminary test of 3-, 4- and 7-class models for organization and location entities showed that the 7-class model outperformed other models. Then, each recognized organization was classified according to its type: 1) University, 2) School, 3) Ministry, 4) Institute, 5) Commercial Company, 6) Centre and 7) Hospital, as well as according to the type of the main descriptor of the organization. The following types of descriptors were used: 1) Chemistry, 2) Biology, 3) Psychology, 4) Health, 5) Medicine/Medical, 6) Pediatric, 7) Surgery, 8) Genetic, 9) Infection, 10) Agriculture, 11) Entomology, 12) Biotechnology, 13) Neurology, 14) Psychology, 15) Pharmacology, 16) Toxicology, 17) Nutrition and 18) Dentistry. An organization belongs to one or another type of descriptor if there is a match between the name of the organization and the name of one of the above descriptors. The organization types and descriptors represent qualitative information and were manually selected based on their observed frequency in the affiliation field. They were mapped to a numeric representation, e.g. types from 1 to 7 and descriptors from 1 to 18.

The Stanford NER was not used for country and city recognition, since the process to identify those entities in such short texts was error-prone. Instead, a dictionary-based method was used. The names of countries and cities were extracted from http://www.geonames.org/. This resource provides a list of city names in different languages. Each city name in the list is mapped to the country name. Thus, we could identify a country associated to the affiliation even in cases where the country name was missing in the affiliation.

2.2 Journal Descriptors and Semantic Types

Frequently, the first author in collaborative publications is the principal contributor in the research work. Other authors can present expertise from different domains. Therefore it is insufficient to measure the similarity of text taken from titles and abstracts for the purpose of AND. To complement this, we used additional descriptors to further define the content of the work. For this purpose a JDI (Journal Descriptor Indexing) tool (Humphrey et al., 2006) developed at the National Library of Medicine (NLM) was used. This tool returns a ranked list of journal descriptors (JD) or UMLS Semantic Types (ST) corresponding to biomedical descriptors as an output to a given text. Ranked items in the output have a score in a range from 0 to 1. There are overall 122 JDs and 135 STs.

Rank	**Score**		**Journal Descriptor**		**Descriptor**	
	PMID 24782557	**PMID** 24481031	**PMID** 24782557	**PMID** 24481031	**PMID** 24782557	**PMID** 24481031
1	0.0178087	0.1916517	JD148	JD148	Pulmonary Medicine	Pulmonary Medicine
2	0.0140019	0.0257541	JD100	JD129	Radiology	Pathology
3	0.0113613	0.0206357	JD023	JD144	Communicable Diseases	Neoplasms

Table 1. Journal Descriptors as output of the JDI tool.

Originally this tool was developed for text categorization purposes with the goal of improving information retrieval. For the AND task an abstract, a title and MeSH terms of articles were provided as an input to JDI. As an example the title, abstract and MeSH terms of the articles with PubMed ID 24481031 and 24782557 were used as input to the JDI tool and the output based on documents counts (Humphrey et al., 2006) is represented either as journal descriptors or semantic types in Tables 1 and 2. In this case the articles were published in the journals "American College of Chest Physicians" and "Respiratory Care," respectively. Both publications share only one MeSH term – "Humans", which is too common and appears in most publications. As it can be seen, the JDs and STs derived from these publications are more descriptive.

Preliminary experiments showed that in most cases the top 3 JDIs have an assigned score much higher than the other JDIs returned. Thus, only the top 3 results were used as an additional feature to describe the domain of a publication.

Rank	**Score**		**UMLS Type**		**Semantic Type**	
	PMID 24782557	**PMID** 24481031	**PMID** 24782557	**PMID** 24481031	**PMID** 24782557	**PMID** 24481031
1	0.5323717	0.6212719	T046	T203	Pathologic Function	Drug Delivery Device
2	0.5264287	0.4946694	T185	T082	Classification	Spatial Concept
3	0.5214509	0.4894958	T169	T046	Functional Concept	Pathologic Function

Table 2. Semantic Types as output of the JDI tool.

2.3 Supervised classifiers

We transform the AND problem to a binary classification task in which a classifier predicts whether the authors of two different publications are the same person. For this purpose, four well-known supervised algorithms (SVM, Random Forest, k-NN and J48) were used to do the classification as well as to evaluate the impact of the features based on Journal Descriptor and Semantic Type to the overall disambiguation performance. These classification algorithms are frequently used in data mining and text-mining tasks (Fernández-Delgado et al., 2014). They have also been used by (Han et al., 2004; Treeratpituk and Giles, 2009; M. Song et al., 2015) for the AND task. The J48 algorithm is a java implementation of the C4.5 algorithm (Quinlan, 2014). All features were normalized according to range of 0 to 1.

2.3.1 *Similarity pairs*

An author profile is represented as an array of values extracted from MEDLINE (name, affiliation, year of publication, etc.), journal descriptors and semantic types. Author profiles are grouped by namespaces. For each namespace, the profiles are compared in a pairwise manner, so that each pair of profiles is represented as a vector of similarity scores between the two profiles. Table 3 shows the process used to transform the discrete values of two profiles into a numeric similarity vector. A Jaro-Winkler algorithm was used to calculate similarity scores for first names of authors. The choice of this algorithm can be explained by its good performance on short strings (M. Song et al., 2015). We chose the SoftTFIDF Jaro-Winkler method to calculate a similarity score for the organizations due to its better performance on longer strings and the fact that it is a less time-consuming algorithm (Cohen et al., 2003). Finally, organization type and journal descriptor were mapped to numeric values and the difference between them was used in the similarity vector.

Profile Values	**Similarity Vector Features**
Full First Name	Jaro-Winkler score ($Full\_First\_Name_a$,$Full\_First\_Name_b$)
Initials	Boolean score ($Initials_a$, $Inititals_b$)
Co-Authors	# of shared co-author names
MeSH terms	# of shared MeSH terms
JDI (3 entities)	# of shared descriptors or semantic types
City	"1" ($City_a = City_b$), "0" ($City_a \neq City_b$)
Country	"1" ($Country_a = Country_b$), "0" ($Country_a \neq Country_b$)
Language	"1" ($Lang_a = Lang_b$), "0" ($Lang_a \neq Lang_b$)
Year	$\lvert Year_a - Year_b \rvert$
Organisation	SoftTFIDF Jaro-Winkler score ($Organisation_a$,$Organisation_b$)
Email	"1" ($email_A = email_B$), "0" ($email_A \neq email_B$)
Type and Descriptor of Organisation	diff ($Type_a$ $Descriptor_a$, $Type_b$, $Descriptor_b$)

Table 3. Similarity vector used to compare the profiles of two authors *a* and *b*.

3 Data

To evaluate the classifiers, a curated corpus for author name disambiguation was used (M. Song et al., 2015). The dataset contains 2,875 publications authored by 385 first authors with 431 author name variants. In less than half of the publications information about emails is present. Furthermore, the majority of the names are of Western origin. Each author in the list has a unique ID assigned by the dataset providers. To date, this is the only known dataset for AND in MEDLINE which is manually curated.

Since the original dataset only consist of author names, PubMed IDs and author IDs, it was necessary to extract all additional relevant information from the MEDLINE corpus. Our final dataset is based on the 2014 MEDLINE/PubMed Baseline Database Distribution. Because the authors considered are only first authors, affiliations are available for the majority of them.

There are articles in 5 different languages in the dataset (denoting the main language of the article's full text, not of the abstract): English, Japanese, Chinese, German and French. The earliest publications are dated from 1967 and the most recent from 2013.

After transformation of pairs of author profiles to similarity vectors, less than a quarter of them belonged to the positive class, i.e. they correspond to the same authors.

4 Results

In this section we present results for each classifier using 10-fold cross-validation. Further, we provide the results of the classifiers from (M. Song et al., 2015) for comparison. Then, we show evaluation scores for the features used in the disambiguation process to rank them according to their contribution.

4.1 Classifier performance

Tables 4-7 show the results obtained by the classifiers. These results are based on three models used to train the classifiers with the following features: (1) Medline features and journal descriptors (MF+JD) obtained with the JDI tool, (2) MF and semantic types (MF+ST) and (3) MF, JD and ST (MF+JD+ST). Additionally we provide the results of the Named Entity Recognizer-based model (NER-based model) and Baseline model described in (Song, Kim et al. 2015). Song's NER model is based on MEDLINE features such as author name, co-authors, affiliation and keywords extracted from article title and journal title. Additionally, Song's NER model relied on the output of the Stanford NER algorithm, which identified organizations, locations and emails in the affiliation text.

Thus, detected entities were transformed into features. Song's Baseline model (M. Song et al., 2015) is based on first author name, article title, and publication venue.

Metrics	MF+JD	MF+ST	MF+JD+ST	NER-Based	Baseline
Precision	0.986	0.975	0.987	0.9776	0.8348
Recall	0.992	0.961	0.994	0.9545	0.8501
F-Measure	0.989	0.9675	0.990	0.9657	0.8423

Table 4. Results of the J48 classifier.

Metrics	MF+JD	MF+ST	MF+JD+ST	NER-Based	Baseline
Precision	0.9785	0.9785	0.991	0.9884	0.8349
Recall	0.9685	0.9725	0.996	0.9634	0.8499
F-Measure	0.973	0.978	0.993	0.9756	0.8322

Table 5. Results of the Random Forest classifier.

Metrics	MF+JD	MF+ST	MF+JD+ST	NER-Based	Baseline
Precision	0.985	0.956	0.987	0.9723	0.8253
Recall	0.988	0.951	0.977	0.9595	0.8412
F-Measure	0.986	0.9535	0.982	0.9656	0.8330

Table 6. Results of the k-NN classifier.

The results achieved on the MF+JD+ST model show a recall which is slightly higher than the precision. In the results of the NER-model the precision has a tiny prevalence over the recall. In Table 7, the precision achieved on models MF+ST and MF+JD+ST is a little greater than the recall, though it is the opposite for the model MF+JD.

Metrics	MF+JD	MF+ST	MF+JD+ST	NER-Based	Baseline
Precision	0.964	0.949	0.9695	0.9541	0.8353
Recall	0.991	0.894	0.905	0.8385	0.8478
F-Measure	0.977	0.9185	0.9335	0.8870	0.8414

Table 7. Results of the SVM with RBF kernel.

4.2 Features Contribution

The information gain feature provided by WEKA was used in order to compute the value of a feature attribute by measuring the information gain with respect to the class. The ranked list of features and their impact according to the information gain is shown in Tables 8 and 9.

The value of the information gain of semantic types is less than that of journal descriptors; see Table 8 and Table 9. In both tables the value of the co-author, year and MeSH-terms features are almost equivalent.

Profile Features	Rank	Value
Full First Name	1	0.310439
Organisation	2	0.292023
Email	3	0.214672
JDIs	4	0.202067
Type and Descriptor of Organization	5	0.183693
Co-Authors	6	0.152621
City	7	0.022
Initials	8	0.01097
Year	9	0.010227
Language	10	0.000584
Country	11	0.000532
MeSH Terms	12	0.0

Table 8. Ranked list of the information gain of features with respect to the class in the MF+JD model.

Profile Features	Rank	Value
Organization	1	0.35203
Full First Name	2	0.287596
Email	3	0.255492
Type and Descriptor of Organization	4	0.20955
Co-Authors	5	0.154428
Semantic Types	6	0.119648
City	7	0.034587
Year	8	0.010847
Initials	9	0.009418
Country	10	0.006007
Language	11	0.000532
MeSH Terms	12	0.0

Table 9. Ranked list of the information gain of features with respect to the class in the MF+ST model.

5 Discussions

The evaluation was performed on the dataset in three different ways: (1) MF+JD, (2) MF+ST and (3) MF+JD+ST. Moreover, we have compared the results to the ones obtained by (M. Song et al., 2015) on the Baseline and NER-based models. Our evaluation results show that the classifiers J48 and Random Forest performed better than the rest. Random Forest provided slightly better results, but in terms of time it was slower than J48. This can be explained by the number of training trees used in the process. The comparison of overall results to Song's NER-Model shows that a significant difference in scores is achieved by our SVM algorithm. However, compared to other classifiers, SVM is less efficient for the AND task and most time-consuming. These results could be explained by the low dimensionality of our data, since SVM performs better on highly dimensional data

The results show also that the MF+JD+ST model outperformed the other models using features based on the topics or descriptors rather than on the keywords or MeSH terms. Nevertheless, the results of the MF+ST model differ from those of the MF+JD. Despite the assumption that the semantic description of the publication would better represent the content, the semantic types and the model MF+ST did not add significant improvement over the MF+JD results. Possible reasons for these results include the fact that the results of the model MF+JD are already sufficiently good, and also that semantic types offer a better characterization of papers than keywords. Surprisingly, the MeSH terms, according to the feature estimation results, showed no impact on the disambiguation model. The information gain of feature attributes also shows that even though the ST-based feature

has a higher impact compared to year, language and location, it only brings slight improvements to the classification results based on the results from the MF+JD+ST model.

The assumption that the location of the author can help disambiguate two profiles was not confirmed. It is not rare when authors change their affiliation and consequently their location. However, in cases where the location of two profiles is identical it suggests that these profiles share the same authorship. An email address, nonetheless, is more significant than a location. The evaluation of features surprisingly showed that journal descriptors and topics are as useful as email addresses for the disambiguation process. Considering that information about the email address of the author is often missing, then the feature based on the journal descriptor and topics can still be used to disambiguate an ambiguous author name.

6 Conclusion

In this paper we have introduced new disambiguation features such as journal descriptors and semantic types, which were not previously used for Author Name Disambiguation. Classification was done with the four most used classifiers for the AND task. The achieved results were compared to state-of-the-art results and it was shown that journal descriptors are as helpful in the disambiguation process as email addresses. Regarding the unexpected value of the MeSH Terms for the classification, the impact of the semantic types to the model can be explained by their nature. Unlike MeSH terms, they are automatically generated for each articles and their granularity is greater.

It is worth mentioning that the results of the evaluation are achieved on the so-called gold standard dataset provided by (M. Song et al., 2015). To date, this is the only dataset which is manually verified. One of the disadvantages of this dataset is that it consists of only first authors of publications. Consequently, the results may be less competitive if datasets consists not only of first authors but also of co-authors. Indeed, in MEDLINE the information about affiliation of co-authors is frequently missing. Moreover, the dataset is biased towards Western types of last names, e.g. Smith, Cohen, Taylor. However, the statistics of most frequent author names in MEDLINE show that they are of Asian origin, for example Wang, Zhang, etc. If we consider that, in the 1990 edition of the Guinness Book of World Records, Zhang was the most common last name in the world, then the disambiguation of certain Asian author names seems extremely challenging. Thus, the classifier models trained on the gold-standard dataset are not necessarily applicable to the AND task for the entire MEDLINE, where non-first authors have missing affiliation and most frequent names are ethnicity-sensitive to the name-matching process (Treeratpituk and Giles, 2012; Jimenez-Contreras et al., 2002; Louppe et al., 2015; Kim and Cho, 2013).

Reference

Cohen, W., Ravikumar, P., & Fienberg, S. (2003). *A comparison of string metrics for matching names and records.* Paper presented at the Kdd workshop on data cleaning and object consolidation.

Cota, R. G., Ferreira, A. A., Nascimento, C., Gonçalves, M. A., & Laender, A. H. (2010). An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. *Journal of the American Society for Information Science and Technology, 61*(9), 1853-1870.

Dlugolinský, Š., Ciglan, M., & Laclavík, M. (2013). *Evaluation of named entity recognition tools on microposts.* Paper presented at the 2013 IEEE 17th International Conference on Intelligent Engineering Systems (INES).

Dogan, R. I., Murray, G. C., Névéol, A., & Lu, Z. (2009). Understanding PubMed® user search behavior through log analysis. *Database, 2009*, bap018.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems. *J. Mach. Learn. Res, 15*(1), 3133-3181.

Finkel, J. R., Grenager, T., & Manning, C. (2005). *Incorporating non-local information into information extraction systems by gibbs sampling.* Paper presented at the Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics.

Haak, L. L., Fenner, M., Paglione, L., Pentz, E., & Ratner, H. (2012). ORCID: a system to uniquely identify researchers. *Learned Publishing, 25*(4), 259-264.

Han, H., Giles, L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). *Two supervised learning approaches for name disambiguation in author citations.* Paper presented at the Digital Libraries, 2004. Proceedings of the 2004 joint ACM/IEEE conference on.

Humphrey, S. M., Lu, C. J., Rogers, W. J., & Browne, A. C. (2006). *Journal descriptor indexing tool for categorizing text according to discipline or semantic type.* Paper presented at the AMIA Annual Symposium Proceedings.

Jimenez-Contreras, E., Ruiz-Pérez, R., & Delgado-Lopez-Cozar, E. (2002). Spanish personal name variations in national and international biomedical databases: implications for information retrieval and bibliometric studies. *Journal Medical Library Association, 90*(4).

Kim, S., & Cho, S. (2013). Characteristics of Korean personal names. *Journal of the American Society for Information Science and Technology, 64*(1), 86-95.

Li, S., Cong, G., & Miao, C. (2012). *Author name disambiguation using a new categorical distribution similarity.* Paper presented at the Joint European Conference on Machine Learning and Knowledge Discovery in Databases.

Liu, W., Islamaj Doğan, R., Kim, S., Comeau, D. C., Kim, W., Yeganova, L., et al. (2014). Author name disambiguation for PubMed. *Journal of the Association for Information Science and Technology, 65*(4), 765-781.

Louppe, G., Al-Natsheh, H., Susik, M., & Maguire, E. (2015). Ethnicity sensitive author disambiguation using semi-supervised learning. *arXiv preprint arXiv:1508.07744*.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*: Elsevier.

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual review of information science and technology, 43*(1), 1-43.

Song, M., Kim, E. H.-J., & Kim, H. J. (2015). Exploring author name disambiguation on PubMed-scale. *Journal of Informetrics, 9*(4), 924-941.

Song, Y., Huang, J., Councill, I. G., Li, J., & Giles, C. L. (2007). *Efficient topic-based unsupervised name disambiguation.* Paper presented at the Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries.

Torvik, V. I., & Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data (TKDD), 3*(3), 11.

Torvik, V. I., Weeber, M., Swanson, D. R., & Smalheiser, N. R. (2005). A probabilistic similarity metric for Medline records: A model for author name disambiguation. *Journal of the American Society for information science and technology, 56*(2), 140-158.

Treeratpituk, P., & Giles, C. L. (2009). *Disambiguating authors in academic publications using random forests.* Paper presented at the Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries.

Treeratpituk, P., & Giles, C. L. (2012). *Name-Ethnicity Classification and Ethnicity-Sensitive Name Matching.* Paper presented at the AAAI.

Warner, S. (2010). Author identifiers in scholarly repositories. *arXiv preprint arXiv:1003.1345*.