Journal of Informetrics 9 (2015) 924–941

Contents lists available at ScienceDirect

Journal of Informetrics

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

Exploring author name disambiguation on PubMed-scale

Min Song*, Erin Hea-Jin Kim, Ha Jin Kim

Department of Library and Information Science, Yonsei University, 50 Yonsei-Ro, Sinchon-Dong, Seodaemun-Gu, Seoul 120-749, South Korea

ARTICLE INFO

Article history:
Received 14 February 2015
Received in revised form 14 August 2015
Accepted 14 August 2015
Available online 24 October 2015

Keywords:
Author name disambiguation
Named entity recognition
Keyphrase extraction
Machine learning
PubMed

ABSTRACT

Author name disambiguation (AND) creates a daunting challenge in that disambiguation techniques often draw false conclusions when applied to incomplete or incorrect publication data. It becomes a more critical issue in the biomedical domain where PubMed articles are written by a wide range of researchers internationally. To tackle this issue, we create a carefully hand-crafted training set drawn from the entire PubMed collection by going through multiple iterations. We assess the quality of our training set by comparing it with SCOPUS-based training set. In addition, for the performance enhancement of the AND techniques, we propose a new set of publication features extracted by text mining techniques. The results of the experiments show that all four supervised learning techniques (Random Forest, C4.5, KNN, and SVM) with the new publication features (called NER model) achieve improved performance over the baseline and hybrid edit distance model.

1. Introduction

The amount of bio-medical data available in PubMed grows exponentially over time, and this kind of big data can support new discovery. Searching for relevant publications from this large literature corpora is a frequent and important task for biomedical researchers and biologists. One common way to search the literature is by author name (Doğan, Murray, Névéol, & Lu, 2009). However, the precision and recall of these searches are typically poor due to ambiguous authors. Indeed, a recent study estimated that about two-thirds of authors in PubMed have ambiguous names (Torvik & Smalheiser, 2009). There are different types of name ambiguity that can confound search results. For example, the same author may appear under distinct names. Alternatively, distinct authors may have matching names. These problems stem from a lack of standards and common practices, and the decentralized generation of content, among other reasons.

Author name disambiguation (AND) is a challenging issue. Disambiguation algorithms may draw false conclusions when faced with incomplete metadata. It becomes a more severe problem in the biomedical domain where PubMed articles are written by a wide range of researchers internationally. There are many techniques proposed for AND and minimizing disambiguation errors (Culotta, Kanani, Hall, Wick, & McCallum, 2007; Liu et al., 2014; Torvik & Smalheiser, 2009). However, those systems were evaluated based on incomplete or incorrect training set (Ferreira, Silva, Gonçalves, Veloso, & Laender, 2012b; Levin, Krawczyk, Bethard, & Jurafsky et al., 2012; Liu et al., 2014) rather than building more accurate evaluation sets due to the bulk of time and efforts required for manually constructing the complete set. Even if there were several attempts to construct the training set manually, automatic mapping during the matching process of author name and publication (Kang, Kim, Lee, Jung, & You, 2011; Levin et al., 2012) causes inaccuracies in the set. Because the AND corpus was built by

* Corresponding author. Tel.: +82 2 2123 2405.
E-mail addresses: min.song@yonsei.ac.kr (M. Song), erin.hj.kim@yonsei.ac.kr (E.H.-J. Kim), hajin_228@yonsei.ac.kr (H.J. Kim).

http://dx.doi.org/10.1016/j.joi.2015.08.004
1751-1577/

such automatic sampling techniques, based on machine learning, the corpus is limited in a sense that it contains inaccurate information.

To tackle this problem, we build the training set from the entire PubMed collection with a multi-step process. The multi-step process involves iteratively and manually collecting and confirming the training set. To our best knowledge, it is the first highly qualified training set created on the PubMed-scale. In addition, throughout the comprehensive literature review on AND, we realize that previous AND studies use the different combinations of publication features and the reported performance varies according to the features used. We measure the impact of various features on the performance of AND techniques and explore how new publication features extracted by text mining techniques impact performance. We conduct a series of experiments of how the performance of several the state-of-the-arts supervised learning techniques changes according to the variety of publication features.

The rest of the paper is organized as follows: The related work section discusses some related work in author name disambiguation. In the section of building AND corpus, we describe how to build the training dataset. The evaluation section reports and discusses the experiment results. In the conclusion section, we summarize key points of our work and present future directions.

2. Related work

Research on AND falls into two categories: one focuses on using AND to extract features of author profiles (Kanani & McCallum, 2007; Koppel, Schler, & Argamon, 2009; Tang, Zhang, Yao, & Li, 2008a; Tang et al., 2008b; Tu, Johri, Roth, & Hockenmaier, 2010); the other focuses on enhancing the performance of AND by implementing machine learning techniques (Cota, Ferreira, Nascimento, Gonçalves, & Laender, 2010; Han, Giles, Zha, Li, & Tsioutsiouliklis, 2004).

There is a huge discrepancy in retrieval performance according to whether the system disambiguates authors or not (Strotmann & Zhao, 2012; Liu et al., 2014). Supervised learning methods that learn features extracted from datasets show better performance in distinguishing identical authors than unsupervised methods, such as clustering techniques (Tang, Yao, Zhang, & Zhang, 2010; Zhang, Tang, Li, & Wang, 2007). However, unsupervised approaches are more favorably adopted due to the drawbacks of supervised learning: (1) building evaluation sets is difficult and time-consuming, and (2) bias of the constructed model with publication features of the restricted datasets due to author productivity distribution (Veloso, Ferreira, Gonçalves, Laender, & Meira, 2012).

For AND research, various publication features have been used, most commonly coauthors, paper title, and publication venue (Cota et al., 2010; Han et al., 2004; Han, Xu, Zha, & Giles, 2005a; Han, Zha, & Giles, 2005b). These three features distinguish a brand new author who has never before appeared in digital collections by heuristic methods (de Carvalho, Ferreira, Laender, & Gonçalves, 2011). Self-citation information is also a good candidate feature to identify authors (Liu et al., 2014; McRae-Spencer & Shadbolt, 2006; Treeratpituk and Giles, 2009).

2.1. Supervised learning approaches

In general, a supervised machine learning approach aims to build a predictive model to distinguish identical authors. To achieve this goal, most supervised learning techniques extract features from publications and learn the extracted features. Zhang et al. (2007) suggest a constraint-based probabilistic model based on six constraints (coaffiliation, coauthor names, citation, coemails, τ-coauthor, and user feedback) between two papers for semi-supervised name disambiguation. The proposed approach shows higher precision (79%), recall (71%), and F-measure (75%) in comparison to the baseline method using the unsupervised hierarchical clustering algorithm. Han et al. (2004) utilize two supervised models: the Naive Bayes (NB) probability model and Support Vector Machines (SVMs). They build the model with the publication list from the researcher's homepage and DBLP databases. SVMs outperforms NB in the collection of webpages (accuracy = 95.6%), while NB shows better accuracy in the DBLP dataset (accuracy = 69.1%).

To determine the best feature selection, Treeratpituk and Giles (2009) use the random forests algorithm combining a collection of decision trees. They extract affiliations, concept-MeSH terms, and so on, in addition to coauthors and journal title from the metadata provided by MEDLINE and compute a wide-range of similarity profiles for 91 selected authors. They found that the most effective variables for disambiguation are Inverse Document Frequency (IDF) of last name and the middle name's similarity. The best disambiguation performance (90% accuracy) is shown with the four variables which include the two variables mentioned above plus affiliations' tfidf similarity and the difference in publication years.

To tackle the limitation of sampling techniques and scalability of AND approaches, Veloso et al. (2012) introduce three levels of associative author name disambiguators: EAND (Eager Associative Name Disambiguation), LAND (Lazy Associative Name Disambiguation), and SLAND (Self-training LAND). They attempted to reduce labeling efforts and find unlabeled authors automatically in the test collection. The results show a micro-averaging F-score 83.3%, compared with 71.2% (NB) and 74.3% (SVM). Ferreira, Veloso, Gonçalves, and Laender (2010) develop a hybrid disambiguation method called SAND (self-training Associative Name Disambiguator). SAND consists of two steps. First, a clustering method is performed to collect examples of training data. Second, supervised disambiguation finds a new author who is not assigned to any given categories in the training examples. As a succession of Veloso et al.'s study (2012), Ferreira et al. (2012b) introduce an active associative sampling method based on association rules in attempt to reduce labeling cost. The results demonstrate the cost-effectiveness of up to 71% in the volume of training datasets.

Several studies attempt to reduce false rates. Culotta et al. (2007) consider the number of publications and coauthor list as first-order features. The training algorithm they proposed is error-driven and rank-based. They reduce errors up to 60% over the common binary classification approach. Wang et al. (2012) propose a four-step boosted-trees method to predict a false paper of same author name. They deal with name and affiliation filtering and construct similarity score in the first and second steps before author screening and boosted trees classification. The proposed boosted trees show a mean of 0.75% in misclassification error but the method remains limited in that a false paper very similar to true paper cannot be detected.

2.2. Unsupervised learning approaches

Unsupervised learning approaches partition author name datasets into publication blocks containing the same first initial and last name. In each block, agglomerative clustering methods are performed based on pairwise similarity in a bottom-up fashion; publications assigned in the same cluster are regarded to be written by the same author (e.g., Cota et al., 2010; Song, Huang, Councill, Li, & Giles, 2007). Assorted techniques are employed such as hierarchical naive Bayes (Han et al., 2005a), k-way spectral clustering (Han et al., 2005b), maximum-likelihood clustering (Torvik & Smalheiser, 2009), bootstrapping method (Levin et al., 2012), heuristic-based hierarchical clustering (Cota et al., 2010), and Latent Dirichlet Allocation (Song et al., 2007; Shu, Long, & Meng, 2009). These methods benefit unlabeled datasets and evaluate their clustering performance by human-annotating of the selected clusters (Liu et al., 2014; Torvik & Smalheiser, 2009) or automatically generated training datasets (Levin et al., 2012; Torvik & Smalheiser, 2009).

Han et al. (2005a) present a hierarchical naive Bayes mixture model based on 14 name datasets extracted from the DBLP database bibliography and two name datasets from author homepages. In their disambiguation method, author's publication vectors are constructed by strings from three key features – coauthor names, paper tile, and journal or proceedings title – because they assume these features contain author's subject area information. The mixture model they suggested outperforms the *K*-means clustering algorithm based on feature vector space model (54.1% to 63.2%). With the same DBLP name datasets, Han et al. (2005b) examine the effects of dataset size on name disambiguation in their previous study (Han et al., 2005a). Liu et al. (2014) lessen false-positive pairing by transitivity violation correction to attain high precision. Their agglomerative clustering algorithm is controlled by name compatibility and probability level when the clusters are merged. As a result, the overall error rate is reduced by 2.0% compared with Authority 2009 (Torvik & Smalheiser, 2009). They note that the unavailability of key PubMed data impairs the disambiguation performance. Therefore, Train prediction models in machine learning methods, researchers must build gold standard publication data sets. Because constructing training datasets is highly labored, several studies create automated positive and negative examples from a whole dataset (Liu et al., 2014; Levin et al., 2012; Torvik, Weeber, Swanson, & Smalheiser, 2005; Torvik & Smalheiser, 2009). Torvik and Smalheiser (2009) generate automated training sets. A positive training set refers to "match set" containing multidimensional comparison vectors sharing the same author name based on computed similarity profiles, while a negative training set (nonmatch set) refers to a different author name's vectors. The model developed by Torvik and Smalheiser (2009) is the extended model that estimates the probability to be the same individual who appears on two different publications based on author-journal similarity profiles (Torvik et al., 2005). Song et al. (2007) extend hierarchical Bayesian text models to calculate topic distributions for name. They use the topic-name probability as feature sets into a hierarchical agglomerative clustering method and measure the similarity between two names by edit distance. The modified agglomerative clustering method outperforms a common clustering method with a mean pair-level pairwise F1 score of 0.911.

3. Building and corpus

As argued by previous studies (Tang et al., 2010; Zhang et al., 2007), the carefully hand-crafted training set for AND is prerequisite to promote research on author name ambiguity. To make the training set most useful, it should be created from large-scale publication collections such as PubMed. In addition, it must embrace both western and Asian names in a harmonious manner. To this end, we employ a multi-step approach. Namely, the AND training set is built in the following three steps:

(1) Author name collection from PubMed
(2) Data cleaning for identifying author name groups
(3) Iterative, manual correction for corpus construction

The AND corpus constructed by this multi-step process is publicly available at https://github.com/Yonsei-TSMM/author_name_disambiguation.git.

Fig. 1 shows how those three steps work. These three-steps are further explained in the subsequent sections.

3.1. Step 1: Author name collection from PubMed

The PubMed dataset that we use for building the training set is a PubMed License 2013 version, which comprises 101.2GB of data (http://www.nlm.nih.gov/bsd/licensee/2014_stats/baseline_med_filecount.html). It consists of 22 million records and 3.5 million author names, typically consisting of a first initial and last name. Because there are many names that belong to

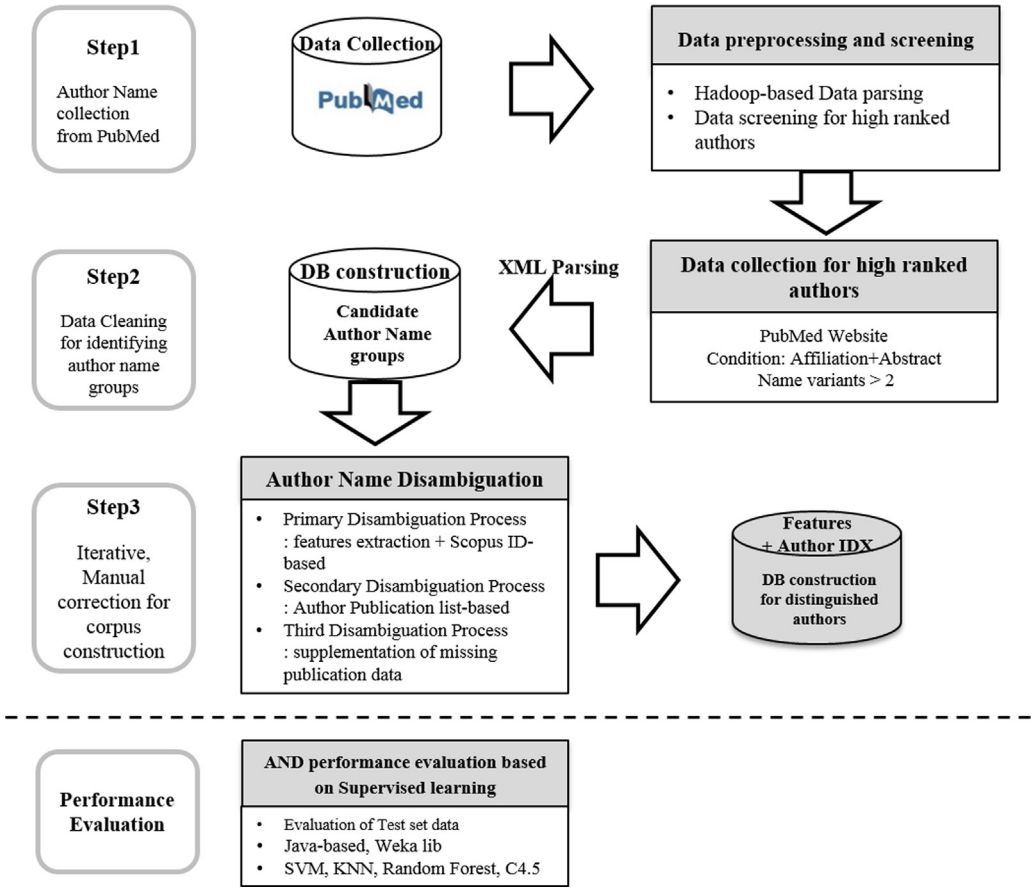


Fig. 1. The process of author name disambiguation.

Table 1
The statistics of PubMed publications.

Publication level	Publications or Groups
Total number of publications	22,000,000
Total number of author name groups with two or more publications	2,071,081
Total number of author name groups with more than 100 publications	14,404

more than one individual, authors are necessarily grouped into author name groups. Within an author name group, there are a number of researchers who share the same name. The number of author name groups who published two or more articles in PubMed is 2,071,081. The number of author name groups who published more than 100 articles is 14,404 (Table 1).

Data Preprocessing and Screening: The PubMed records we downloaded from the NLM site (http://www.nlm.nih.gov/databases/journal.html) are stored in XML, and we developed a Hadoop-based XML parsing module due to the ample size of the dataset. Hadoop, developed by Apache (http://hadoop.apache.org/), is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers. It reduced the processing time to 1/5 that of a single-node computer. In addition, we calculated the frequency of author names with Hadoop. The list of author names is sorted by frequency, and the author name consists of "first initial + last name." The reason that we do not use "first full name" is because many of PubMed records provide first initial only, and thus we have to unify the author names. Fig. 2 shows an example of author name groups by frequency. Finally, we extract author name groups with more than 540 publications which results in 670 author name groups.

3.2. Step 2: Data cleaning for identifying author name groups

Among highly ranked 670 author name groups, we select top 67 identified by Hadoop-based calculation of author name frequency. The author name group is chosen for an author who bears a minimum of 500 publications containing both organization and abstract in XML data; the organization information is helpful to cluster the same authors manually at the beginning stage, while abstract gives the subject information to distinguish authors sharing the same name at the final stage. We collect

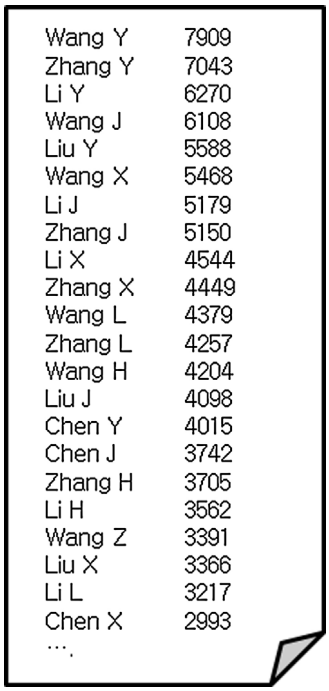


Fig. 2. The top N author name groups.

all of PubMed XML records written by those selected authors from PubMed (http://www.ncbi.nlm.nih.gov/pubmed/). The total number of collected records is 26,715. From these records, we extract features for AND. We utilize these features to disambiguate the selected authors. Here is the list of features we used: first author last name, first author initials, affiliation, PMID, article title, coauthor list, journal title, and abstract.

Among these features, extracting author affiliation and the keyword list requires further attention for processing. Because only partial author affiliation is available in each record (primarily the first author's affiliation is available), we need a way of identifying coauthors' affiliation information either manually or programmatically. In addition, there is discrepancy in author affiliation depending on how an author enters his or her affiliation information and journal style. Thus, we decide to apply a Named Entity Recognition technique to extract organization and country from affiliation entered by authors (Manning et al., 2014). This way, we can minimize the string mismatch that may exist in affiliations of the same author. For the keyword list, because a record is not always accompanied by author-provided keywords, we apply the keyword extraction technique, called MAUI, to extract keywords from the title and the abstract.

The coauthor list is an important indicator of who works with whom, and can be a prominent feature for AND (Zhang et al., 2007; Han et al., 2004; Treeratpituk and Giles, 2009; Veloso et al., 2012; Wang et al., 2012). MeSH terms are also an interesting feature because the indexer assigned MeSH terms to records and common MeSH terms shared by authors imply that they have the similar research interests (Liu et al., 2014). Because our training set is collected from the entire PubMed collection, however, too many records do not contain MeSH terms, and we decided to exclude them in the publication feature set for the performance evaluation.

3.3. Step 3: Iterative, manual correction for corpus construction

An AND corpus can be constructed more accurately with the training set of selected authors by performing the following three processes, illustrated in Fig. 3.

3.3.1. Primary disambiguation process

As previously mentioned, we select the top-ranked 67 author name groups. The total number of the records published by those author name groups is 26,715. In the step of Primary Disambiguation Process, we first identify those 67 author name groups based on his or her affiliation information, emails, and coauthor list. This identification process requires more clues to distinguish distinct authors because the affiliation information is incomplete. Thus, we decided to use an external source, SCOPUS, to assist our labored disambiguation by assigning the SCOPUS Author Identifier (SCOPUS ID) from SCOPUS Database (www.scopus.com) as an Author Identifier. The SCOPUS ID is a unique single identifier that is assigned to each author by the automatic grouping algorithm in SCOPUS based on various meta data such as author's affiliation, coauthors, and subject area, etc. (http://help.scopus.com/Content/h_autsrch_intro.htm). We used PMID as a query to retrieve

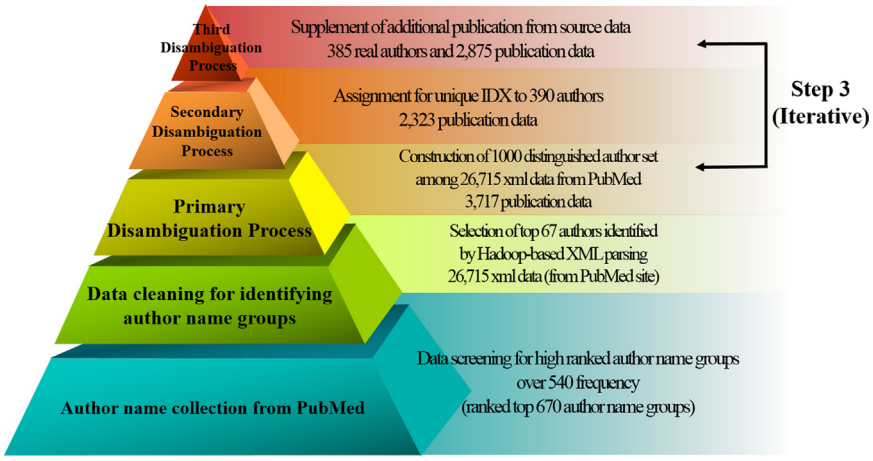


Fig. 3. The construction of the evaluation set.

records from SCOPUS Database and extracted SCOPUS IDs that match PMIDs from the retrieved results. By doing this, we are able to collect the list of pairs of SCOPUS ID and PMID. These procedures led to identify 67 author name groups based on his or her affiliation information, emails, coauthor list, and the external resource, SCOPUS ID. As a result, we refine 1000 author sets among 26,715 records, for a total of 3717 publications. These 1000 author sets are available for download at https://github.com/Yonsei-TSMM/author_name_disambiguation.git. Among these 1000 author sets, several publications of researchers are omitted when using SCOPUS ID, and in many cases, a SCOPUS ID matches authors incorrectly.

3.3.2. *Secondary disambiguation process*

To disambiguate the 1000 selected author set from the first step, we manually verify the authors' publication lists from their website and conduct AND. We search the authors' website to clarify author names more precisely.

This second step represents how we find publication lists in the authors' websites. First, we consider the feature set that is constructed through data pre-processing as a search query. Second, we combine and input the keywords related to authors to search for websites to gather the authors' publication lists. For instance, we use a Google query (Kang et al., 2011) to retrieve the website containing the publication list (i.e., Google query = intitle: Joachim Cohen), or we use a Boolean operator as a query (i.e., Google query = Jordi Cohen and University of Illinois). The authors' publication lists are regarded as the source data drawn from the website containing the article lists. Through this procedure, we identified that top-ranked retrieval results are the author profile lists that were provided by professional online network services such as BioMedExperts (http://www.biomedexperts.com) and Labome (http://www.labome.org/platform.html). Since author lists provided by the professional online network services often do not contain the appropriate publication list of the authors, we have to manually search for authors' publication lists on the Web, which take a substantial amount of time to retrieve and confirm the author's correct publication list on the Web. This manual identification step enables us to confirm or reject the initial identical author identification of the 67 author name groups.

Applying the aforementioned querying method, we finally assign a unique ID to 390 authors who are selected from 1000 authors, and the total number of publications is 2323. Those 390 authors include both their own publication lists and source data. Although the various types of websites provide authors' publication lists, most of the professional author service sites such as ResearchGate (http://www.researchgate.net), BioMedExperts, and Labome give us the publication lists based on automatic extraction algorithms, which suffer from lack of reliability. Therefore, we only collect the publication lists from websites provided by authors themselves, and use their affiliations as a source for author name disambiguation.

3.3.3. *Third disambiguation process*

To build a higher-quality training set, we supplement additional publications found in the author publication list to our author dataset and ultimately inspect the distinguished author sets. By adding the omitted data to the originally gathered set of PMIDs, we increase the total number of the publications. Identifying the distinguished authors who are selected manually from the Primary Process to the Third Process, we finally construct the AND training set, which contains 385 authors and 2875 publications. In other words, the final AND training set, a gold standard author set, has 385 authors among 36 groups of authors, and 2875 publications.

Table 2 shows the characteristics of the source data from publication features that were constructed from PubMed. We cover both western and eastern author names in a harmonious manner. In addition, the training set embraces enough name variants per author name group to evaluate the performance of AND techniques. For example, "Waston R" has three types of different name variants which represent "Waston R," "Waston Roger," and "Waston Rebecca." Having diverse name variants improves our training set for AND. In the final 36 author name groups, 431 name variants exist.

Table 2
The author statistics of the PubMed evaluation set.

Author name	Name variants	Records	Author name	Name variants	Records
Agarwal, R	9	284	Martin, C	5	42
Anderson, C	8	34	Miller, M	8	63
Banerjee, S	18	110	Moore, A	2	12
Brown, J	28	124	Nakamura, H	13	159
Cohen, J	7	59	Park, J	14	66
Evans, M	25	92	Patel, S	18	142
Gardner, J	3	14	Romano, M	6	41
Ghosh, S	26	168	Roy, S	21	122
Gupta, R	14	88	Schmidt, H	6	61
Jain, S	27	145	Smith, R	9	73
Johnson, D	2	8	Sun, X	16	71
Jones, R	5	38	Taylor, J	22	111
Kaiser, J	2	37	Thomas, L	9	29
Khan, M	10	37	Markman, M	2	78
Klein, R	2	33	Watson, R	3	54
Lee, J	51	252	Weber, M	10	50
Liu, F	11	86	Williams, N	4	16
Lutz, S	3	11	Zhang, D	12	65

Appendix A shows the example of an identified Author "Miller M," one of the author names from the AND corpus. The names are assigned IDX, which represent distinguished real author, and the validation source shows the URL of the publication list that we manually searched on the Web through Primary and Secondary stage. Those who have same validation source with same IDX in the corpus represent a unique distinguished author.

4. The quality and the limitations of the training set

To evaluate the quality of our training set, we build the comparative training set with SCOPUS data. To this end, we collect SCOPUS author IDs and publication data by retrieving publication with PubMed IDs in our training set. The final SCOPUS training set misses 374 out of 2875 publications (12.94%) of our training set. This significant rate of missing publication indicates that our training set contains the comprehensive publication list. For 2503 publications excluding the missing publications, we manually compare our training set with the SCOPUS training set. We identified that 170 out of 2503 publications (6.79%) have incorrect author IDs assigned. For example, for the distinct author name, "ML Evans," the SCOPUS training set assigns the same ID "23967774500" to the different authors that happen to share the same name. Another case is that multiple author IDs are assigned to one author. For example, for author "S Ghosh," three SCOPUS author IDs are assigned. This is because the author "S Ghosh" is indexed with three different names: "AR Ghosh," "SK Ghosh," and "S Ghosh." Unless a manual, iterative confirmation process is used, it is difficult to obtain a highly accurate training set like ours, which in turn helps evaluate the AND technique more accurately.

There are 3,654,892 author name groups by first name in PubMed, 1,597,337 (43.7%) of which have only one publication. That leaves 2,057,555 author name groups that have published in PubMed at least twice, with their names appearing on a total of 20,402,663 publications. Author name groups that appear in PubMed more than 100 times account for just 0.7% of author name groups appearing at least two times, but the number of the papers written by these highly prolific author name groups is 3,062,437 (about 15%). We select the top 670 author name groups to include as the candidates for identifying the first author. These author name groups appear more than 540 times in PubMed and account for about 0.03% of total PubMed authors, but about 3.5% of PubMed publications. We examine the last names of these author name groups and identify that 182 distinct names exist in them. Because there are Western names predominately included in these author name groups, we decide to include Eastern name groups even if they are lower ranked. Finally, we select 36 author name groups that represent about 19.7% of 182 author name groups.

Building a highly accurate, comprehensive training set is a daunting challenge. Although we initially select 670 author name groups, we end up including 36 of them due to the fact that including all of 670 author name groups requires the tremendous amount of time. This is the major limitation of our training set. However, unlike previous training sets, our training set is 100% human curated and consists of a relatively larger number of author name groups and publications. Previous training sets were constructed either by (1) group clustering of random sample followed by post manual confirmation by evaluators (Liu et al., 2014) or (2) publication URLs selected in a semi-automatic fashion (Kang et al., 2011). In other domains, the commercial training set constructed by author name disambiguation techniques is used as the training set (Torvik & Smalheiser, 2009). In addition, existing human curated training sets consist of very little data. Han et al. (2004) manually created the training set that consists of only 15 different "J Anderson" authors with 219 publications. Our 36 author name groups include 385 authors and 2875 publications.

One downside of using our training set for the purpose of AND is that even if we include both Eastern and Western names, the majority (about 70%) of the training set is still Western names. Thus, it may not be adequate for AND research that focuses on Eastern author name disambiguation. Second, since the training set consists mainly of highly productive authors, for AND

research that focuses on less productive authors, our training set may not be suitable in that the publication features of less productive authors are different from ones of highly productive authors. Third, the training set is built by using the first authors as IDX. Thus, coauthors may not be well represented in the training set. Although it is not abnormal to only consider the first authors as identifiers (Zhang et al., 2007; Song et al., 2007), we will address this limitation in a future version of the training set. Lastly, our training set does not address the problem of the same authors who have different names, which is a challenging task of author name disambiguation.

5. Experiments

There are two main evaluation approaches to author name disambiguation: author grouping and author assignment (Ferreira, Gonçalves, & Laender, 2012a). The author grouping method uses clustering techniques to group authors by similarity functions such as Cosine or Jaccard (Han et al., 2005b) whereas the author assignment method uses either supervised learning or model based clustering techniques to predict the best match of a publication to an author with the author model with the labeled train corpus (Treeratpituk and Giles, 2009). The proposed approach belongs to the author assignment method. We built an author model by learning whether publication A and B are written by the same author. With the author model, we predict whether publication C and D are written by the same author or not. In other words, we predict whether publications they have authored belong to the same "class." Several previous studies adopted this pair-wise classification (Ferreira et al., 2010; Han et al., 2004; Torvik et al., 2005; Treeratpituk and Giles, 2009).

We conduct a series of experiments to investigate whether and how publication features have an impact on resolving name disambiguation in PubMed. In addition to the publication features adopted by most previous AND studies, we employ state-of-the-art text mining techniques such as the named entity recognition (NER) technique and the keyword extraction technique. The necessity of applying NER to AND stems from the fact that, in PubMed records, an affiliation field contains all information about author's organization, country, and e-mail address, and the pattern of organization information displayed is diverse. We detect location, organization, and e-mail addresses using the NER technique provided in Stanford CoreNLP (Manning et al., 2014). In addition, because author- or indexer-provided keywords are not always available, it is difficult to judge the topical similarity between two authors. To tackle this issue, we use the MAUI keyphrase extraction technique to extract keywords from the title and the abstract (Medelyan & Witten, 2008). We measure the performance of three different combinations of publication features, such as the vector space similarity model (we call the baseline), the hybrid edit distance model that combines the vector space similarity and edit distance, and the NER-based model.

5.1. Classifier

To evaluate the performance the aforementioned three models, we use four well-accepted supervised learning algorithms in the study of AND, provided in WEKA (http://www.cs.waikato.ac.nz/ml/weka/).

The four demonstrated classifiers in our AND experiments are:

(1) Random Forest: an ensemble supervised learning algorithm that construct multiple decision trees at training time to avoid the decision tree's problem of over-fitting to their training set.
(2) KNN (k-Nearest Neighbor): a supervised learning algorithm that stores all available cases and classifies new cases based on a similarity measure.
(3) C4.5: a type of decision tree algorithm that builds decision trees from a set of training data using the concept of information entropy.
(4) SVMs (Support Vector Machines): a discriminative learning algorithm that outputs an optimal hyperplane that categorizes new examples given labeled training data.

We set parameters for these four classifiers implemented in WEKA as follows: For Random Forest, we set the number of trees in the Random Forest algorithm to 10. For each tree, a bootstrap sample of the same size is created to serve as the training data. Only a random subset of the available features of defined size is considered for each node. WEKA's Random Forest is not based on CART or J48; rather it is based on a variant of REPTree that is modified to include the desired randomness. Thus, we use the default values for sample size and node size set by WEKA API. For SVM, we set the complexity parameter (−C) to 0.1, which SVM uses to build the hyperplane between any two target classes. We set the gamma parameter to 1.0, which is critical for classification performance. We use the default values for the other parameters, which the API document says are not critical for performance.

Pair-wise similarity metrics: A publication is represented as a collection of publication features. The pair-wise algorithm calculates the similarity scores between the corresponding features of any two publications by using similarity metrics. In the present paper, we adopt two popular similarity metrics, Jaccard's and Jaro–Winkler distance.

Table 3
Feature set used as the input for feature similarities.

	IDX	First author last name	First author initial	Affiliation	Coauthors	Paper title	Journal title
Pub_A	3	Smith	R	Department of Computer Science, Brigham Young University, Provo, Utah, USA. 2@gmail.com	Williamson, Ryan; Ventura, Dan; Prince, John T	Rubabel: wrapping open Babel with Ruby	Journal of cheminformatics
Pub_B	3	Smith	R	Department of Computer Science, Brigham Young University, Provo, UT 84602, USA. 2aaa@gmail.com	Ventura, Dan	A general model for continuous noninvasive pulmonary artery pressure estimation	Computers in biology and medicine

The Jaccard's similarity metric (JSM), also called the Jaccard's distance function, is used to estimate the similarity between two vectors (or attributes). It is suitable to calculate the similarity for a long vector such as a paper or a journal title. The Jaccard's similarity score of two objects (i.e., publications) X and Y, JSM (X,Y), is calculated as follows:

$$\mathrm{JSM}(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{1}$$

where $X \cap Y$ is the number of same strings in X and Y, and $X \cup Y$ is the total number of strings in X and Y.

For example, suppose that there are two journal titles such that X = "Computers in biology and medicine" and Y = "Computational biology and chemistry." $X \cap Y$ is 2 since terms "biology", "and" are common, and $X \cup Y$ is 7 since the following 7 terms, "computer", "in", "biology", "and", "medicine", "computational", "chemistry", appear in X and/or Y. Thus, the Jaccard's similarity score, JSM (X,Y), is 2/7 which is 0.29.

The Jaro–Winkler distance algorithm measures the distance between given two string, $s1$ and $s2$. In the Jaro distance J_d (Jaro, 1995) formula below, $|s1|$ is the length of the first string and $|s2|$ is the length of the second string. Let string $s1 = a_1 \ldots a_m$ be the characters in the $s1$ and let string $s2 = b_1 \ldots b_k$ be in the characters in $s2$. m refers to the number of same characters in $s1$ and $s2$. t is half the number of transpositions and the transpositions define when the character of position i has different sequence order in $s1$ and $s2$ such that $a_i \neq b_i$ (Cohen, Ravikumar, & Fienberg, 2003). Jaro distance J_d formula is:

$$J_d = \begin{cases} 0 & \text{if} \quad m = 0 \\ \frac{1}{3}\left(\frac{m}{|s1|} + \frac{m}{|s2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases} \tag{2}$$

Jaro–Winkler distance (Winkler, 1999) JW is:

$$\mathrm{JW} = J_d + (l \times p(1 - J_d)) \tag{3}$$

where l is the length of common prefix up to maximum of 4 characters, and p is a common value for the constant in Winkler, which is $p = 0.1$. For example, given the string $s1$ "MANUEL" and $s2$ "MANEUL", we find $m = 6$, $|s1| = 6$, $|s2| = 6$. There are mismatched characters U/E and E/U, which present $t = 2/2 = 1$. Jaro distance (J_d) can be $J_d = 1/3\left((6/6) + (6/6) + ((6-1)/6)\right) = 0.944$. For Jaro–Winkler distance, we find $p = 0.1$, $l = 3$, $J_d = 0.9444$, and the Jaro–Winkler distance JW is 0.961 ($0.9444 + (3 \times 0.1(1 - 0.944))$).

We use the Jaccard's similarity metric to calculate the similarity of the pair of features between two publications except for first author name. Jaro–Winkler distance works better on the shorter string.

We built disambiguation models learned from selected features with the training set for each classifier. In total, twelve cases (the baseline, hybrid model, and NER-based model by four classifiers) are compared.

Input for the classifier: The proposed approach was built on top of WEKA. Thus, the input for the classifiers we implement follows the WEKA format for classification, which is called the ARFF (Attribute-Relation File Format) format consisting of relation, attribute, and data. The conversion procedure of the raw bibliographic records to the ARFF format goes through the following steps:

(1) We first extract six features from bibliographic records. An author identification number is assigned to each record. Table 3 shows the extracted features that are used as an input for computing feature similarities between two publications (Smith & Ventura, 2013a; Smith et al. 2013b).
(2) In the second step, we build an author model to learn whether two publications are written by the same author or not by pair-wise similarity. The example input for pair-wise similarity is shown below.

Table 4
The example of the ARFF format.

Header	@relation FirstAuthorName
	@attribute 1_ x_{1AB} numeric
	@attribute 2_ x_{2AB} numeric
	@attribute class {0,1}
Data	@data
	0.8,0.9,1
	0.2,0.9,0
	0.7,0.7,0

- Pub_A = {first_author_last_name$_A$, first_author_initial$_A$, affiliation$_A$, coauthors$_A$, paper_title$_A$, journal_title$_A$}
- Pub_B = {first_author_last_name$_B$, first_author_initial$_B$, affiliation$_B$, coauthors$_B$, paper_title$_B$, journal_title$_B$}

With Pub_A and Pub_B, the pair-wise similarity of a feature set between two publications is computed using Jaccard's similarity in the following manner:

SimMetrics x_{AB} = {x_{1AB}, x_{2AB}, x_{3AB}, x_{4AB}, x_{5AB}, x_{6AB}} where x_{1AB} is the first_author_last_name feature of Pub_A and Pub_B.

With the case of Table 3, the SimMetrics value of first author's initial and last name by Jaccard's algorithm is {1,1} for of PubA, PubB because the first author last name of the pair (Smith) is the same and the first author initial (*R*) is the same.

These SimMetrics values are converted to the ARFF format as the input for the classifier. An ARFF file is an ASCII text file containing a list of instances sharing a set of attributes. An ARFF file consists of two distinct sections: header information and data information (http://www.cs.waikato.ac.nz/ml/weka/arff.html).

Suppose that we have three instances where (1) SimMetrics1 = {0.8, 0.9} for PubA and PubB labeled as the same author, and it is classified to class 1, (2) SimMetrics2 = {0.2, 0.9} for PubA and PubC labeled as the different author that is classified to class 0, and (3) SimMetrics3 = {0.7, 0.7} for PubB and PubC labeled as the different author that is classified to class 1. With these instances, the ARFF format is generated as shown in Table 4.

(3) In the third step, we predict whether two publications are written by the same author with the trained author model. Note that the same conversion process of the training phase above is applied to the prediction phase.

5.2. Publication feature selection

Table 5 shows the summary of previous AND studies and the set of publication features used in each.

Table 5 presents publication features used for AND in previous studies. Selected features are classified according to the method of name disambiguation: supervised and unsupervised learning approaches.

Many previous studies have drawn features from metadata of publications. In supervised learning approaches, most studies exploit the coauthors names, paper title, and journal titles as features (Han et al., 2004; Treeratpituk and Giles, 2009; Veloso et al., 2012; Wang et al., 2012). Also, features are calculated by the similarity vector between two authors' publications (Culotta et al., 2007; Treeratpituk and Giles, 2009). Some of the features such as co-affiliations and co-emails are defined by the relation between two papers and used for a publication feature (Zhang et al., 2007). Wang et al. (2012) further consider authors' metadata as a feature set.

In unsupervised learning approaches, the common publication features are similar to ones that are used in supervised learning; coauthor names, paper title and journal title (Han et al. 2005a; Liu et al., 2014; Cota et al., 2010). However, Liu et al. (2014) further exploit features such as MeSH term (keyword), grant, and publisher names. Levin et al. (2012) also extend the set of features, which include subject category, middle initials, last name initials, reprint organization, and citing and cited features.

We select various important features, including the most common features such as first author, coauthor names, and affiliation, as well as unique features introduced in this paper such as entities extracted by the named entity recognition technique and keywords extracted by the keyword extraction technique. As shown in Table 5, few previous studies used the NER and key extraction technique for the feature set on author name disambiguation.

5.3. Evaluation measure

To measure the performance of AND techniques, we perform 10-fold cross validation to estimate a classifier's accuracy should that classifier be constructed from all of the training data. We use WEKA API, which is a well-accepted routine for k-fold cross validation in supervised learning tasks. K-fold cross validation divides the original data set into *k* subsets of equal size. K-1 subsets are used for constructing the model, while the remaining subset is used to test the classifier. This process is repeated *k* times, and the average of validation results is reported (Kohavi, 1995). In our experiments, the training set of 2875 publications (by 385 authors in 36 author name groups) is randomly partitioned into 10 subsets, nine of which are used to build the binary classification model. The remaining subset is used for validating the model. We repeat this validation 10 times and compute the average of 10 validation results.

Table 5
Publication features from the previous related works.

	References	Paper-level	Author-level
Supervised Learning Approaches	Han et al. (2004)	(1) Coauthor names (2) Paper title (3) Journal title	
	Zhang et al. (2007)	(1) Coauthor names (2) CoAffiliation (3) Coemails (4) Citation (5) User Feedback (6) τ-coauthor	
	Culotta et al. (2007)	(1) First and middle names of the author (2) Number of overlapping coauthors (3) Title of the two publications (4) Author emails (5) Affiliations (6) Venue of publication	
	Treeratpituk and Giles (2009)	(1) Coauthors (2) Authors (3) Affiliations (4) Journal title (5) Concept-MeSH Term (6) Article title	
	Veloso et al. (2012)	(1) Coauthors names (2) Journal title (3) Publication venue	
	Wang et al. (2012)	(1) Coauthor names (2) Paper title (3) Abstract words (4) Keywords (5) Subject category (6) Number of authors (7) Number of authors' affiliations (8) Cited Journals	(1) Number of papers (2) Author's research field (3) Asian surname (4) Surname commonness
Unsupervised Learning pproaches	Han et al. (2005a)	(1) Coauthor names (2) Paper title (3) Journal title	
	Liu et al. (2014)	(1) Coauthor names (2) Affiliation (3) Paper title (4) Abstract words (5) Journal title (6) MeSH (7) Publisher names (8) Publication year (9) Grant (10) Substance	
	Cota et al. (2010)	(1) Coauthor names (2) Paper title (3) Publication venue	
	Levin et al. (2012)	(1) Publication year (2) Subject category (3) Last Names with Initials (4) Addresses (5) Middle Initials (6) E-mail Address (7) Language (8) Reprint Organization (9) Citing and Cited Feature	

For performance measure, we adopt pairwise precision, recall, and *F*-measure that are computed based on the confusion matrix. These pairwise measures are variations of precision, recall, and *F*-measure that are well-accepted measures for any machine learning papers (Li, Cong, & Miao, 2012). In our evaluation, the performance of classification algorithms is evaluated in a binary manner: any two publications are predicted to belong to the same class.

$$\text{Pairwise} \quad \text{Precision} = \frac{\text{the number of pairs correctly predicted}}{\text{the total number of pairs predicted}} \tag{4}$$

$$\text{Pairwise} \quad \text{Recall} = \frac{\text{the number of pairs correctly predicted}}{\text{the total number of correct pairs}} \tag{5}$$

Table 6
The list of publication features and the similarity measure used in three prediction models.

	Baseline	Hybrid	NER
• Feature sets	• The first author name (first initial and last name) • Paper title and journal (or proceeding) title (connected with whitespace)	• The first author name (first initial and last name) • Coauthor list • Paper title and journal (or proceeding) title (connected with whitespace)	• The first author name (first initial and last name) • Coauthor list • Organization, location, and e-mail detected from affiliation by NER • Keywords extracted from paper title and journal (or proceeding) title (connected with whitespace) by MAUI
• Similarity metric	• Jaccard's	• Jaro–Winkler for first author • Jaccard's for the rest of features	• Jaro–Winkler for first author • Jaccard's for the rest of features

$$\text{Pairwise F-measure} = \frac{2(\text{PP} \cdot \text{PR})}{\text{PP} + \text{PR}} \tag{6}$$

Two papers that are classified into the same label in the gold standard dataset are called a correct pair, and two papers that are predicted to be in the same label in the gold standard dataset by a classifier but should be classified into a different label are called a wrong pair.

Suppose that there are four documents A, B, C, and D. Combinations of the document pair are as follows: (A,B) = same author, (A,C) = same author, (A,D) = different author, (B,C) = same author, (B,D) = different author, and (C,D) = different author. If the classifier predicts that (A,B) is to be the same author, (A,C) the same, (C,D) the same but (A,D), (B,C), and (B,D) not be the same author, pairwise precision (PP), pairwise recall (PR), and pairwise F-measure (PF) are calculated such that:

(1) PP = 4/6 = 0.67
(2) PR = 4/6 = 0.67
(3) PF = $2(0.67 \times 0.67)/(0.67 + 0.67) = 0.67$

where the total number of correct pairs is 6, the number of pairs correctly predicted is 4, and the total number of pairs predicted is 6.

5.4. Prediction model for author name disambiguation

As mentioned earlier, we build three prediction models with the different combinations of publication features and compare their performance. Table 6 shows the summary of these three models and the similarity measure used.

5.4.1. Baseline model

As shown in Table 5, the most common features of the machine learning in AND research are coauthors, paper titles, and publication venue (Torvik & Smalheiser, 2009; Han et al., 2004; Cota et al., 2010). Publication venue is used to imply the topic correlation of the same author (Yang, Peng, Jiang, Lee, & Ho, 2008). Thus, for the baseline, we select first author names, paper titles, and publication venue. We exclude the coauthor list from the baseline features, instead we use the coauthor features in the hybrid and the NER-based models because few previous studies do not use coauthors as a feature (Culotta et al., 2007; Levin et al., 2012). We measure the impact of coauthor feature on the performance.

5.4.2. Hybrid edit distance model

We adopt edit distance in the hybrid edit distance model by incorporating edit distance into calculating similarity between two features, as the second baseline. To this end, we use string similarity instead of token-based binary similarity. As mentioned earlier, because Jaro–Winkler works well in a short array of strings, we apply Jaro–Winkler for the feature of first author name.

5.4.3. NER-based model

In this model, we add distinctive features that previous studies did not use, such as the organization name recognized by NER and the keyword list by MAUI. By NER, we detect organizations, locations, and e-mail addresses in affiliation field from the PubMed dataset. The author's affiliation takes various forms in PubMed data (i.e., "University of Hull," "School of Nursing, University of Hull, Hull, UK. xxx@hull.ac.uk," or "School of Nursing, Social Work and Applied Health Studies, University of Hull, HU6 7RX, Hull, UK. xxx@hull.ac.uk"). We utilize the NER technique available in Stanford CoreNLP to find out the correct organization information and e-mail address.

We first compute similarity vectors based on selected features and similarity coefficient between publication pairs. The machine learning algorithms are applied to similarity vectors and create predictive models in the learning phase. We then

predict whether two authors are identical in an author pair using the predictive models built in the learning phase. We evaluate the prediction accuracy based on answer set, i.e., labeled dataset that we have constructed in evaluation phase.

6. Results

6.1. Baseline and hybrid edit distance model

Figs. 4 and 5 present the disambiguation performance on the four different classifiers. The baseline is shown in Fig. 4 and the Hybrid edit distance model is in Fig. 5. The first author name, paper title, and publication venue are selected in the baseline. The similarity of publication pairs is calculated with token-based similarity. In Fig. 4, we can see consistent recalls, precisions, and F-measures (82% to 85%) in all four classifiers.

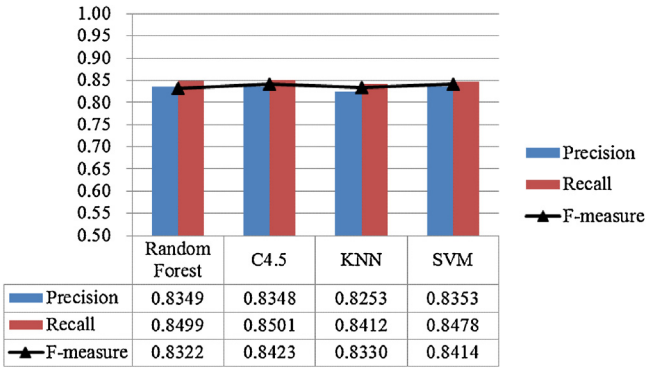
The precision increases when the coauthor list is added to the publication feature set in the Hybrid edit distance model (Fig. 5). Random Forest, KNN, and C4.5 achieve 95.95%, 95.91%, and 96.09%, respectively. However, this results in only a very slight increase in the case of SVM (83.53% to 83.81%). The experiment results indicate that the coauthor feature provides the distinctive power to determine identical authors because adding the coauthor list outperforms the baseline in Random Forest, KNN, and C4.5, but we need more distinctive features that can capture uniqueness to an author class for SVMs.

Fig. 5 shows that the Random Forest classifier 91.94% F-measure, 95.95% precision outperforms KNN, C4.5, and SVM approaches (91.80%, 91.74%, and 84.80% F-measure, respectively) when using coauthor information in the training set.

6.2. NER-based model

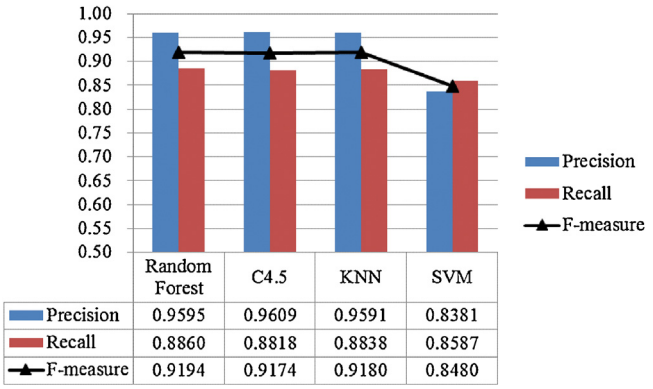
Fig. 6 show the results obtained by the NER-based model. The NER-based model shows the superb performance in all four classifiers compared to baseline and hybrid model. The NER-based model achieves all over 95% (precision, recall, and F-measure) except for recall and F-measure of SVM.

With regard to the performance difference among four classifiers, in F-measure, Random Forest achieves the highest performance (97.56%) and C4.5 is the second highest (96.57%). KNN is 96.56% followed by SVM (88.7%) shown in Table 7 and Fig. 6. For recalls, Random Forest also achieves the highest performance (96.34%) and KNN is the second (95.95%). C4.5 is 95.45% followed by SVM (83.85%). Our suggested model shows the best performance in precision, recall, and F-measure over



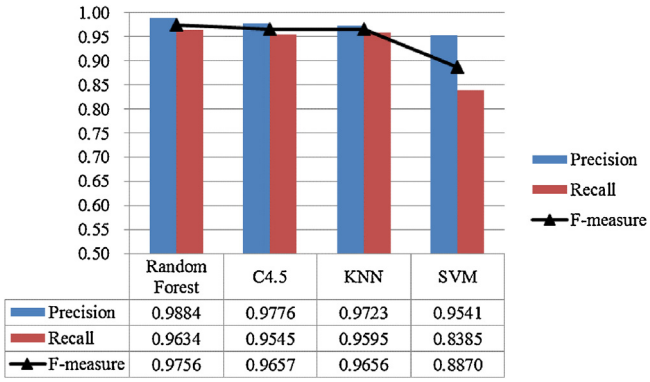
	Random Forest	C4.5	KNN	SVM
Precision	0.8349	0.8348	0.8253	0.8353
Recall	0.8499	0.8501	0.8412	0.8478
F-measure	0.8322	0.8423	0.8330	0.8414

Fig. 4. The results of baseline with the training set.



	Random Forest	C4.5	KNN	SVM
Precision	0.9595	0.9609	0.9591	0.8381
Recall	0.8860	0.8818	0.8838	0.8587
F-measure	0.9194	0.9174	0.9180	0.8480

Fig. 5. The results of the hybrid edit distance model with the training set.



	Random Forest	C4.5	KNN	SVM
Precision	0.9884	0.9776	0.9723	0.9541
Recall	0.9634	0.9545	0.9595	0.8385
F-measure	0.9756	0.9657	0.9656	0.8870

Fig. 6. The results of the NER-based model.

both baseline and the Hybrid model except SVM classifier's performance. Random Forest classifier seems the best choice for the author disambiguation among the four classifiers by our experiments.

We summarize the overall performance of the three models in Table 7. We conducted an analysis-of-variance (ANOVA) to evaluate the statistical significance between the three models. The independent variable, the model factor, included three levels: the baseline, Hybrid, and NER-based models. The NER-based model shows the significant performance over the both baseline and Hybrid models at $p < 0.05$.

6.3. The impact of feature sets on the performance

To examine the impact of different combinations of features on the performance, we conducted three experiments in addition to baseline, hybrid, and NER: (1) B + C denotes baseline + coauthor list, (2) B + C + E denotes baseline + coauthor list + extracted organization, location, and email entities, (3) B + C + E + K denotes baseline + coauthor list + extracted organization, location, and email entities + keyphrase list. Note that baseline consists of first author name (first initial and last name) + paper title and journal title (connected with white space). The difference between the B + C model and the hybrid model is that, in the B + C model, we apply the Jaccard's similarity metric for all features, whereas in the hybrid model we use the Jaro–Winkler algorithm for the first author name and the Jaccard's similarity metric for the rest of the features. These two similarity measures also differentiate the B + C + E + K model from the NER model.

Overall, our experiments indicate that including more features for classification tends to improve the model's performance in F-measure (Fig. 7). Among six models, the NER model achieves the best performance in all three measures. Features of extracted entities such as organization, location, and email confer the most significant improvements. This contradicts previous studies that reported coauthor list as the most influential feature for performance. This discrepancy may be attributed to the fact that the entity feature is unique and different from the affiliation feature used by existing AND research that contains many variations of author's affiliations. In addition, the experiment results indicate that the performance of classification algorithms with keyphrase list is better than one with bag of word from paper and journal titles in precision. However, in terms of recall, the keyphrase list does not significantly improve the performance; in some cases, it even results in inferior

Table 7
The results of overall performance of experiments.

Classifier	Evaluation	Baseline	Hybrid model	NER-based model
Random Forest	Precision	0.8349	0.9595	0.9884
	Recall	0.8499	0.8860	0.9634
	F-measure	0.8322	0.9194	0.9756
C4.5	Precision	0.8348	0.9609	0.9776
	Recall	0.8501	0.8818	0.9545
	F-measure	0.8423	0.9174	0.9657
KNN	Precision	0.8253	0.9591	0.9723
	Recall	0.8412	0.8838	0.9595
	F-measure	0.8330	0.9180	0.9656
SVM	Precision	0.8353	0.8381	0.9541
	Recall	0.8478	0.8587	0.8385
	F-measure	0.8414	0.8480	0.8870

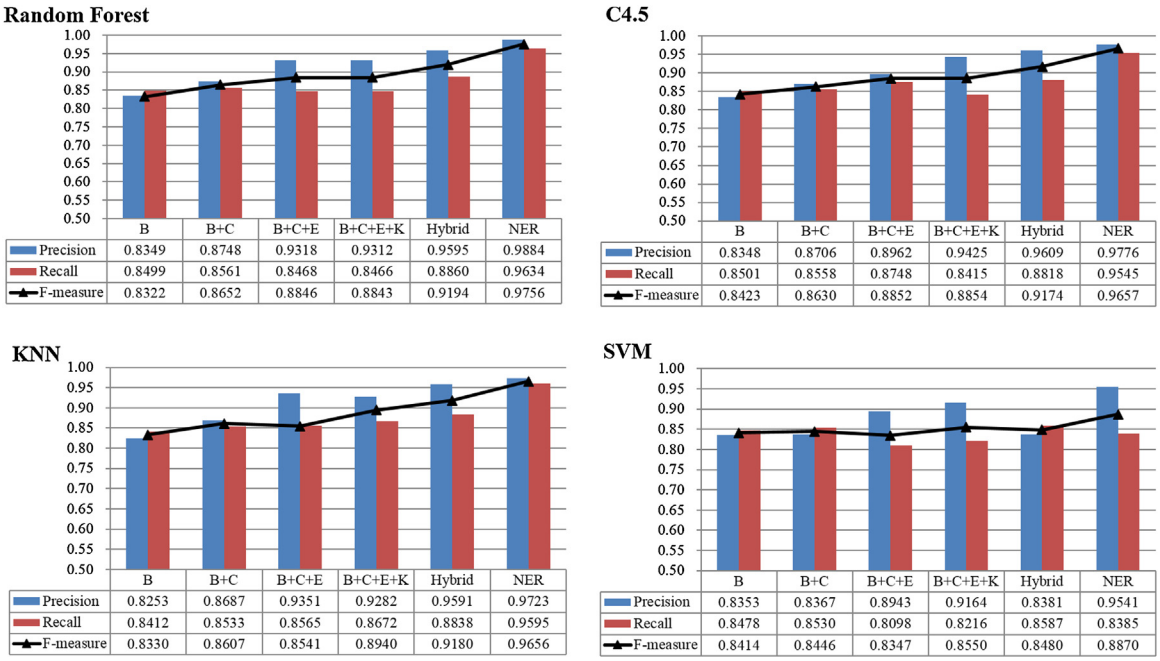


Fig. 7. The impact of different combinations of feature sets on the performance of classifiers.

performance. Unlike the other three algorithms, the performance of SVM shows inconsistent performance. In the case of SVM, the hybrid model achieves the best performance in recall followed by the B + C model. The B model also shows the strong performance. In the analysis-of-variance (ANOVA) test, adding the entity feature and the keyphrase list to the model (B + C + E + K) shows the significant performance improvement over the baseline at $p < 0.05$. However, B + C and B + C + E are not statistically significant.

7. Conclusion

In this paper, we explore PubMed-scale AND techniques that are aided with a carefully hand crafted training set and a wide range of publication features including novel features extracted by text mining techniques. The training set is constructed with three steps of publication data processing. The major contribution of the proposed work is that, to our best knowledge, it is the first AND study that provides a large-scale, highly qualified training set and novel features.

The resulting training set and publication features turn out to be very precise and accurate compared with existing training sets and approaches generated from commercial databases such as SCOPUS. We evaluated our approach in three different ways: (1) baseline model, (2) Hybrid edit distance model, and (3) NER model. The experiments show that the NER model significantly outperformed the other two models using state-of-the-art supervised learning techniques in measures of precision, recall, and F-measure. The statistical test confirms that the NER model shows a significantly better performance than the baseline and Hybrid models at $p < 0.05$.

The limitation of the present work is that although we compare our training set with the SCOPUS training set for the quality measures, we did not make the formal comparison between our training set and the training sets created by previous AND studies. We made our training set publicly available so that other AND studies such as semi-supervised or unsupervised learning approaches can evaluate their algorithms with our training set. We also plan to apply the proposed approach to data collections of other domains such as DBLP and BDBComp. In addition, we plan to build a new training set that is suited to the problem of the same authors who have different names, a challenging task of author name disambiguation.

Acknowledgements

This work was supported by Korea Institute of Science and Technology Information (KISTI) in Republic of Korea (file no. K-14-SG-23-02N-1) and (in part) by the Yonsei University Future-leading Research Initiative of 2014 (2014-22-0116).

Appendix A. An example of an idenfied author Miller, M from constructed author name data set

IDX	Pmid	Journal_Title	Article_Title	Journal _PubDate	First_Author _Last_Name	First_Author_ Fore_Name	First_Author_ Initials	Affiliation	Coauthor	Validation source
280	19584133	American journal	Suicide among...	2009	Miller	Matthew	M	Harvard Injur.	Barber, C...	http://www.hsph.harvard.edu/matthew-miller/publications/
280	20801585	Drug and alcohol	Exposure to alc...	2011	Miller	M	M	Department....	Borges, G...	http://www.hsph.harvard.edu/matthew-miller/publications/
280	21391934	Journal of the Am	Opioid analgesic.	2011	Miller	Matthew	M	Department....	Stürmer, T...	http://www.hsph.harvard.edu/matthew-miller/publications/
280	22224886	Annual review of	Suicide mortalit...	2012	Miller	Matthew	M	Harvard Injur.	Azrael, D...	http://www.hsph.harvard.edu/matthew-miller/publications/
280	22390578	American journal	Preventing suici...	2012	Miller	Matthew	M			http://www.hsph.harvard.edu/hicrc/firearms-research/gun-ownership-and-use/
280	22390591	American journal	Veterans and su...	2012	Miller	Matthew	M	Department....	Barber, C...	http://www.hsph.harvard.edu/matthew-miller/publications/
280	22390593	American journal	A call to link da...	2012	Miller	Matthew	M	Harvard Injur.	Azrael, D...	http://www.hsph.harvard.edu/hicrc/firearms-research/gun-ownership-and-use/
280	23597351	American journal	Method choice...	2013	Miller	Matthew	M	Department....	Hempstead...	http://www.hsph.harvard.edu/matthew-miller/publications/
280	24146116	CNS drugs	Antidepressant...	2014	Miller	Matthew	M	Department....	Pate, V...	http://www.hsph.harvard.edu/matthew-miller/publications/
283	9197549	Gene	Alternative splic...	1997	Miller	M	M	Department....	Zeller, K...	http://medschool.umaryland.edu/facultyresearchprofile/viewprofile.aspx?id=476
283	9581716	Journal of the Am	Normal triglyce....	1998	Miller	M	M	Division of C.	Seidler, A...	http://medschool.umaryland.edu/facultyresearchprofile/viewprofile.aspx?id=476
283	9714130	Arteriosclerosis	Apolipoprotein...	1998	Miller	M	M	Department....	Aiello, D...	http://medschool.umaryland.edu/facultyresearchprofile/viewprofile.aspx?id=476
283	11152845	The American jou	Prevalence of...	2001	Miller	M	M	Division of C.	Rhyne, J...	http://medschool.umaryland.edu/facultyresearchprofile/viewprofile.aspx?id=476
283	16415199	Heart (British Car	Impact of cinem...	2006	Miller	M	M		Mangano, C	http://medschool.umaryland.edu/facultyresearchprofile/viewprofile.aspx?id=476
283	18279736	Journal of the Ame	Impact of trigly...	2008	Miller	Michael	M	Division of C.	Cannon, C...	http://medschool.umaryland.edu/facultyresearchprofile/viewprofile.aspx?id=476
283	20368475	Psychosomatic m	Divergent effect...	2010	Miller	Michael	M	Division of C.	Mangano, C	http://medschool.umaryland.edu/facultyresearchprofile/viewprofile.aspx?id=476
283	20713894	Circulation	Hold the patty,...	2010	Miller	Michael	M			http://medschool.umaryland.edu/facultyresearchprofile/viewprofile.aspx?id=476

References

Apache. (2007).;1; Retrieved from http://hadoop.apache.org/.

BioMedExperts. (2013).;1; Retrieved from http://www.biomedexperts.com.

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.

Cohen, W. W., Ravikumar, P., & Fienberg, S. E. (2003). A comparison of string distance metrics for name-matching tasks. In *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03)* Acapulco, Mexico,

Culotta, A., Kanani, P., Hall, R., Wick, M., & McCallum, A. (2007). Author disambiguation using error-driven machine learning with a ranking loss function. In U. Nambiar, & Z. Nie (Eds.), *Sixth international workshop on information integration on the web (IIWeb-07)* Vancouver, Canada.

de Carvalho, A. P., Ferreira, A. A., Laender, A. H., & Gonçalves, M. A. (2011). Incremental unsupervised name disambiguation in cleaned digital libraries. *Journal of Information and Data Management*, *2*(3), 289.

Doğan, R. I., Murray, G. C., Névéol, A., & Lu, Z. (2009). Understanding PubMed user search behavior through log analysis 2009. *Database: The Journal of Biological Databases and Curation*, *2009*, bap018.

Ferreira, A. A., Veloso, A., Gonçalves, M. A., & Laender, A. H. (2010). Effective self-training author name disambiguation in scholarly digital libraries. In J. Hunter, C. Lagoze, L. Giles, & Y. F. Li (Eds.), *Proceedings of the 10th annual joint conference on digital libraries* (pp. 39–48). Gold Coast, Australia: ACM.

Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. (2012). A brief survey of automatic methods for author name disambiguation. *ACM Sigmod Record*, *41*(2), 15–26.

Ferreira, A. A., Silva, R., Gonçalves, M. A., Veloso, A., & Laender, A. H. (2012). Active associative sampling for author name disambiguation. In K. B. Boughida, B. Howard, & M. L. Nelson (Eds.), *Proceedings of the 12th ACM/IEEE-CS joint conference on digital libraries* (pp. 175–184). Washington, DC, USA: ACM.

Han, H., Giles, L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In H. Chen, H. Wactlar, & C. Chen (Eds.), *Proceedings of the 2004 Joint ACM/IEEE conference on digital libraries* (pp. 296–305). Tucson, USA: IEEE.

Han, H., Xu, W., Zha, H., & Giles, C. L. (2005). A hierarchical naive Bayes mixture model for name disambiguation in author citations. In H. M. Haddad, & L. M. Liebrock (Eds.), *Proceedings of the 2005 ACM symposium on applied computing* (pp. 1065–1069). Santa Fe, USA: ACM.

Han, H., Zha, H., & Giles, C. L. (2005). Name disambiguation in author citations using a k-way spectral clustering method. In M. Marlino, T. Sumner, & F. Shipman (Eds.), *Proceedings of the fifth ACM/IEEE-CS joint conference on digital libraries* (pp. 334–343). Denver, USA: ACM.

Jaro, M. A. (1995). Probabilistic linkage of large public health data files. *Statistics in Medicine*, *14*(5-7), 491–498.

Kanani, P., & McCallum, A. (2007). Efficient strategies for improving partitioning-based author coreference by incorporating Web pages as graph nodes. In *Proceedings of AAAI 2007 workshop on information integration on the Web* Vancouver, Canada, (pp. 38–43).

Kang, I. S., Kim, P., Lee, S., Jung, H., & You, B. J. (2011). Construction of a large-scale test set for author disambiguation. *Information Processing & Management*, *47*(3), 452–465.

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection*. Quebec, Canada: IJCAI.

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, *60*(1), 9–26.

Labome. (2008). Retrieved from ⟨http://www.labome.org/platform.html⟩

Levin, M., Krawczyk, S., Bethard, S., & Jurafsky, D. (2012). Citation-based bootstrapping for large-scale author disambiguation. *Journal of the American Society for Information Science and Technology*, *63*(5), 1030–1047.

Li, S., Cong, G., & Miao, C. (2012). Author name disambiguation using a new categorical distribution similarity. In *Machine learning and knowledge discovery in databases*. Berlin Heidelberg: Springer.

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.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In D. Marcu, & K. Toutanova (Eds.), *Proceedings of 52nd annual meeting of the association for computational linguistics: System demonstrations* Baltimore, USA, (pp. 55–60).

McRae-Spencer, D. M., & Shadbolt, N. R. (2006). Also by the same author: AKTiveAuthor, a citation graph approach to name disambiguation. In G. Marchionini, M. L. Nelson, & C. C. M. Marshall (Eds.), *Proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries* (pp. 53–54). Chapel Hill, USA: ACM.

Medelyan, O., & Witten, I. H. (2008). Domain independent automatic keyphrase indexing with small training sets. *Journal of American Society for Information Science and Technology*, *59*(7), 1026–1040.

NLM. (2014). Retrieved from http://www.nlm.nih.gov/bsd/licensee/2014_stats/baseline_med_filecount.html.

ResearchGate. (2008). Retrieved from http://www.researchgate.net.

SCOPUS. (2004). Retrieved from www.scopus.com

Shu, L., Long, B., & Meng, W. (2009). A latent topic model for complete entity resolution. In Y. Ioannidis, D. Lee, & R. Ng (Eds.), *IEEE 25th international conference on data engineering* (pp. 880–891). Shanghai, China: IEEE.

Smith, R., & Ventura, D. (2013). A general model for continuous noninvasive pulmonary artery pressure estimation. *Computers in biology and medicine*, *43*(7), 904–913.

Smith, R., Williamson, R., Ventura, D., & Prince, J. T. (2013). Rubabel: Wrapping OpenBabel with Ruby. *Journal of Cheminformatics*, *5*, 35.

Song, Y., Huang, J., Councill, I. G., Li, J., & Giles, C. L. (2007). Efficient topic-based unsupervised name disambiguation. In E. Rasmussen, R. Larson, E. Toms, & S. Sugimoto (Eds.), *Proceedings of the 7th ACM/IEEE-CS joint conference on digital libraries* (pp. 342–351). Vancouver, Canada: ACM.

Strotmann, A., & Zhao, D. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, *63*(9), 1820-1833.

Tang, J., Zhang, J., Yao, L., & Li, J. (2008). Extraction and mining of an academic social network. In J. Huai, R. Chen, H. W. Hon, Y. Liu, & W. Y. Ma (Eds.), *Proceedings of the 17th international conference on World Wide Web* (pp. 1193–1194). Beijing, China: ACM.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). Arnetminer: Extraction and mining of academic social networks. In Y. Li, B. Liu, & S. Sarawagi (Eds.), *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 990–998). Las Vegas, USA: ACM.

Tang, J., Yao, L., Zhang, D., & Zhang, J. (2010). A combination approach to web user profiling. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *5*(1), 2.

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*. In F. Heath, M. Lively, & R. Furuta (Eds.), *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries* Austin, USA: ACM.

Tu, Y., Johri, N., Roth, D., & Hockenmaier, J. (2010). Citation author topic model in expert search. In *Proceedings of the 23rd international conference on computational linguistics: Posters*. Association for Computational Linguistics.

Veloso, A., Ferreira, A. A., Gonçalves, M. A., Laender, A. H., & Meira, W. (2012). Cost-effective ondemand associative author name disambiguation. *Information Processing & Management*, *48*(4), 680–697.

Wang, J., Berzins, K., Hicks, D., Melkers, J., Xiao, F., & Pinheiro, D. (2012). A boosted-trees method for name disambiguation. *Scientometrics*, *93*(2), 391–411.

WEKA. (2009). Retrieved from http://www.cs.waikato.ac.nz/ml/weka/.
Winkler, W. E. (1999). The state of record linkage and current research problems. In *Statistical Research Division*. US Census Bureau.
Yang, K. H., Peng, H. T., Jiang, J. Y., Lee, H. M., & Ho, J. M. (2008). Author name disambiguation for citations using topic and web correlation. In *Research and advanced technology for digital libraries*. Berlin Heidelberg: Springer.
Zhang, D., Tang, J., Li, J., & Wang, K. (2007). A constraint-based probabilistic framework for name disambiguation. In A. H. F. Laender, & A. O. Falcão (Eds.), *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 1019–1022). Lisbon, Portugal: ACM.