Research on Author Name Disambiguation Based on Fusion Features and Semantic Fingerprints

Xiaorui Zhai, Hongqi Han*, Zhong Li and Yaxin Ran
Institute of Scientific and Technical Information of China, Beijing 100038, China; Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content, SAPPRFT
*Corresponding Author. Email: bithhq@163.com

Abstract. Author name disambiguation has been a challenging problem in many applications. In order to promote researches to solve name disambiguation, Aminer launched the Open Academic Data Challenge 2018 jointly with Chinese Association for Artificial Intelligence and China Knowledge Centre for Engineering and Technology. Aminer is a scholar-centered academic search and mining platform covering more than 200 million papers and more than 100 million scholars in various academic fields. Our team proposed a name disambiguation method based on fusion features and semantic fingerprint technique to participate in the competition. The method identified authors with same names through organization feature and co-author feature at first, and then it solves ambiguity names by way of semantic fingerprints which are 128-bit binary vector generated from textual features of papers by Simhash algorithm. Our method scored 0.609 on the verification set and 0.879 on the test set ranking first in the final submission.

1. Introduction
Author name ambiguity usually refers to the phenomenon that several authors share a same name. It is a very common and has been a challenging issue in digital library. For applications based on scientific and technical literatures, e.g. academic collaboration network construction, scientific research ability evaluation, community detection and academic recommendation, it is the first problem to be solved [1, 2]. Many researches reveal that ambiguity author names will influence the effectiveness. However, it is not easy to solve name polysemy problem, and the massive growth of scientific literature has made this problem more difficult and urgent. Although many studies on name disambiguation have been presented in the past years, the problem has not been solved very well.

In response to this situation, the Joint Research Centre of Knowledge and Intelligence of Tsinghua University (K&I Tsinghua) co-sponsored the Open Academic Data Challenge 2018 with China Knowledge Centre for Engineering Sciences and Technology and Chinese Association for Artificial Intelligence. The competition task is to identify which authors of the same name belong to the same person. The sponsor provided competition datasets which contains a large number of challenging authors with the same name in Aminer. Aminer is a scholar-centred academic search and mining platform managed by K&I Tsinghua, and covers more than 200 million papers and more than 100 million scholars in various academic fields. The author's name disambiguation is a core function as well as a challenge problem in Aminer [3].

About 337 teams registered to participate in the competition. Participants are allowed to use open-source code or tools, while any unpublished or unauthorized code or tools are prohibited. All
participants can use pretrained and open-source word embeddings, but are not allowed to use any additional academic data. Each team can train disambiguation algorithm using training dataset, and optimize their model by making submissions for validation dataset. After the test dataset is opened, each team can make final submissions. The top score will be selected to a team for multiple submissions. Our team participate in the competition, and employed a method based on fusion features and semantic fingerprints which is based on our previous research [4]. The score of our method is 0.87 and ranked first in the final submissions.

2. Brief Review of Author Name Disambiguation

The existing author name disambiguation methods can be classified into two categories, namely rule-based disambiguation and machine learning-based disambiguation. The rule-based approach primarily develops effective rules instituted by domain experts and then completes disambiguation according to the rules. Zhu [5] constructed five rules using co-authors, provinces, organizations, article categories and research fields, while Smalheiser and Torvik [6] only used co-authors to set disambiguation rules.

The machine learning-based disambiguation method can be divided into supervised learning, semi-supervised learning and unsupervised learning according to whether the annotated information exists in the dataset used. The supervised learning-based method uses the manually labelled dataset to learn the classification rules and train the classification model, and then performs name disambiguation based on the model. Han et al. [7] built a professional category library by mining Web information, and then used it as training set to complete the disambiguation model training. The unsupervised learning-based method includes cluster-based method and deep learning-based method. The cluster-based method selects the context information or social relationship related to the person's name, and then the names whose similarity comparison results of related information meet the threshold will be regarded as similar name. Common clustering algorithms include hierarchical clustering [8], density clustering [9], and network clustering [10]. The deep learning-based method learns distributed representation from the data, i.e. the text data is represented in the form of word vector, and then the generated word vector is used as the initial input of the neural network model in deep learning. After that, the documents and names to be disambiguated will be predicted to some categories based on the learning of the initial input. Chiu and Nichols [11], Gridach [12], Qamas et al. [13] both used LSTM for disambiguation. The semi-supervised learning method combines the supervised learning method and the unsupervised learning method, using a small amount of manual annotated dataset and lots of unlabelled dataset at the same time, such as [14].

3. Author Name Disambiguation Method

3.1. The Process of Author Name Disambiguation

The proposed disambiguation method is based on three features: organization feature, co-author feature and textual feature. The process of author name disambiguation is shown in figure 1. In the first step, organization feature is extracted to identify same authors, where same name with same organization are clustered into same authors. In the second step, co-author feature is used to identify same authors based on co-authorship and the clusters in the last step, where same names with same co-authors are assigned to same authors. In the last step, textual feature is used to create semantic fingerprint for each article, and name disambiguation is resolved according the comparing results of fingerprints.
3.2. Author Assignment Based on Organization Feature

Organization feature is one of the high recognition characteristics for identifying same name authors [15]. In most cases, two authors with same name can be judged to be same person if their organizations are same. Therefore, we select the organization as one of the disambiguation features. Although there may be multiple forms of expression in the same institution, there are still similarities between different forms of expression. So, we select the edit distance to calculate the similarity of the author institutions. If the edit distance of the organization corresponding to the disambiguated authors in any two papers is greater than threshold $\alpha$, the two papers are considered belonging to the same author. The specific process is shown in figure 2.

3.3. Author Assignment Based on Co-author Feature

The co-author feature is another feature for easily identify same authors with same names. If two articles have two same co-author names, they probably might be written by one individual author. Smallheiser and Torvik [6] mentioned such an example in their study. Considering the good effect in name disambiguation, we choose the co-author feature to assign same name to same individual. Giving the author assignment results based on organization feature, a list of co-authors for each paper in the same category is obtained. For a paper to be assigned (author name of which is to be disambiguated), if the paper has more than two co-authors (including two, except the disambiguated author name) with any paper in a category, the paper is assigned to the category. Figure 3 shows an example of the assignment process for paper 4 to category 1 and category 2. If the number of co-authors is less than two, the paper would be not assigned to any category.
3.4. Author Assignment Based on Semantic Fingerprints

Generally, the assignment according to the organization and co-author features is effective and usually has high precision in author name disambiguation, however the recall is comparatively low [4]. In order to further improve the effect of disambiguation, we select the text feature to optimize the preliminary disambiguation results. Fingerprint technique has been proved effective for duplicate or plagiarism detection at first [16-18], and also has testified to have good performance in author name disambiguation [4].

The text features are used to generate text fingerprints for every paper. One paper with ambiguity name will be assigned to a category created in the two steps if the fingerprint similarity compared with papers in the category is high. However, a paper may be assigned to multiple categories or can't be assigned any category. In the first case, the paper will be decided according to the designed arbitration rule, while the paper will be assigned to a new same-name author in the second case.

3.4.1. Generating Simhash Fingerprints. The text feature for creating fingerprints consists of title, abstract, keyword and venue name. All the textual fields are concatenated into a long string with a space character for each article. Then the strings are used to generate semantic fingerprints with 128-bit binary vector by the Simhash algorithm [19].

3.4.2. Author assignment of unclassified papers. After assignment based on organization and co-author features, we obtain some clusters containing articles written by same authors; however, we may have lots of paper unassigned. The fingerprint of each unassigned paper is compared with every paper in found clusters. The equation (1) based on hamming distance is defined to evaluate the similarity between two papers, where p denotes the similarity result, Len (Simhash) is the length of the Simhash fingerprint (128 in the method), and Hamming Distance is the distance of two fingerprints.

\[
p = \frac{\text{Len(Simhash)} - \text{HammingDisatance}}{\text{Len(Simhash)}}
\]  

(1)

Two thresholds \( \theta_1 \) and \( \theta_2 \) are set in order to determine the assignment of a paper, where \( \theta_1 > \theta_2 \). An unclassified paper will be assigned to a category when p is greater than \( \theta_1 \) comparing the similarity with any paper in the category, while it won't be assigned when p is less than \( \theta_2 \). If p is between \( \theta_1 \) and \( \theta_2 \), the paper will be decided in the arbitration stage. The specific process of arbitration is shown in figure 4. After compared with papers in all categories, the unclassified paper will be assigned to be a new author category, i.e. the paper is written by a new same-name author.
3.4.3. Arbitration. An unclassified paper will be decided who is the real author by means of arbitration if it is assigned to multiple author categories. For these papers, we use a simple comparison of similarity to complete assignment. The average similarity between the paper with all articles in an author category is calculated. Then the paper is assigned to the category with the highest average similarity.

4. Experiments and Discussions

4.1. Dataset Description and Evaluation Metric
The datasets in the competition are categorized as training set, validation set and test set. The training set is used to train disambiguation models for participated teams, while the validation set is used for optimizing their models. The test set is opened for the final score of all teams by submitting their disambiguation results in the last few days in competition. The Macro Pairwise-F1 [20] based on Pairwise Precision; Pairwise Recall is the evaluation metric for submissions, as shown in equation (2). The definitions of Pairwise Precision and Pairwise Recall are shown in [20].

$$\text{PairwiseF1} = \frac{2 \times \text{PairwisePrecision} \times \text{PairwiseRecall}}{\text{PairwisePrecision} + \text{PairwiseRecall}}$$ (2)

4.2. Threshold Selection
The thresholds of the method include the threshold $\alpha$ for judging the similarity of the organization, and the threshold $\theta_1$ and $\theta_2$ for judging the similarity of the text. By analysing the organization fields in given dataset, 0.8 was selected as the value of threshold $\alpha$. We randomly selected 10 authors in the test set, compared the results of textual similarity between any two papers of one author, and analysed the comparison results between any two documents of same author or different authors. As a result, the range of $\theta_1$ was set between (0.50, 0.60), and the range of $\theta_2$ was set between (0.60, 0.70).

Since there were too much noise and incomplete fields in the dataset, how to optimize thresholds is a big problem in the competition. Although the evaluation index is provided by the sponsor, however the evaluation codes were not open for every team, we couldn't get same evaluation score for our submissions when adopting the given equation of evaluation index. Thus, we could only optimize the thresholds by multiply submitting the predicted results of the validation sets. Finally, $\theta_1$ and $\theta_2$ were set to 0.70 and 0.55, respectively.
4.3. Experimental Results and Evaluation

For every submission, the competition system just gave a score without precision and recall. The highest score is taken as the final result for multiple submissions. In order to know which parameters achieved best result, we recorded thresholds corresponding to each submission. The best competition score of our method and its parameters are shown in Table 1.

| Experiment Method | Thresholds           | Dataset and scores |
|-------------------|----------------------|--------------------|
|                   |                      | Verification set   |
| Method based on Organization and Co-author | \( \alpha = 0.8 \) | 0.574              |
| Method based on Fusion Features and Semantic Fingerprints | \( \alpha = 0.8, \theta_1 = 0.55, \theta_2 = 0.70 \) | 0.609              |

As a whole, we adopted two strategies trying to achieve good disambiguation results. One strategy is that we utilized fusion features of organization and co-author for achieving high precision; however the score based on the verification set was 0.574 because the recall is low. Another strategy is that we used comparisons between semantic fingerprints of same-name articles to improve precision and recall so as to achieve higher F-measure, and the score based on the verification set was 0.609. After the test set was released, the score for the method based on organization and co-author features was 0.534 and the score for the method based on fusion features and semantic fingerprints was 0.879.

It can be seen from our submission that the disambiguation results based on proposed method is obviously better than those based only on the organization and co-author feature. The disambiguation method based on single feature, such as organization, co-author and text, is likely prone to lower recall or precision, while the proposed method can exploit the advantage of every single feature and implement better effect.

5. Conclusions

Author name disambiguation is usually the fundamental step in scientific literature management, character search, and social network analysis. The article presents an author name disambiguation method for the Open Academic Data Challenge 2018. The method is based on fusion features and semantic fingerprints. The organization feature and co-author feature are fused to disambiguate same names preliminarily, and then the method uses text feature to resolve name ambiguity by comparing the similarity of semantic fingerprints which are generated by Simhash algorithm. In the final submission, 35 teams submitted their disambiguation results. The score of the presented method is 0.879 which ranks first among all teams.

Although our study has achieved good performance on the competition dataset, there are some issues need to be paid attention to. The first issue is large amount of noise data in the training set might influence the performance of other models in the competition. The second issue is large amount of missing or incomplete abstract might bring less trouble for our method than other text-based method. The third issue is parameter training and optimization might be misled for our method and other machine learning method because author assignment results of verification set are not open during competition.

In the future, we will improve our method in the following aspects: (1) we will optimize parameters in the process of threshold selection for achieving better disambiguation effect. (2) The algorithm of semantic fingerprint used in the article just exploits the textual information in corpus. The semantic information is "local" which may influence author assignment effect because the fingerprint of an article is generated according to its own textual data. Next, we will try to use a "global" algorithm to train semantic fingerprint based on large corpus in advance, and will compare the effects of two methods.
6. Acknowledgments
This work is mainly supported by the National Natural Science Foundation of China (Project 71473237), and partially supported by The ISTIC Key Works Project (ZD2018-07) and The Program of the China Knowledge Centre for Engineering Science and Technology (CKCEST-2018-1-26). The authors are grateful to the National Natural Science Foundation of China, the Ministry of Science and Technology of China, and the Chinese Academy of Engineering for their financial support to carry out this work.

7. References
[1] Yuan J P, Yu Z L, Su C, Ma Z, Yang Z Q and Su J 2011 A survey of author name disambiguation Digit. Libr. Forum 10 60-65
[2] Fu Y, Zhu L J and Han H Q 2016 A survey of name disambiguation Technol. Intell. Eng. 2 53-58
[3] https://www.biendata.com/competition/scholar2018/
[4] Han H Q, Yao C Q, Fu Y, Yu Y S, Zhang Y L and Xu S 2017 Semantic fingerprints-based author name disambiguation in Chinese documents Scientometrics 111 1879–1896
[5] Zhu Y X 2014 Study on author name disambiguation for Chinese bibliographic information Libr. and Inf. Serv. 58 143-148
[6] Smalheiser N R and Torvik V I 2009 Author name disambiguation Annu. Rev. of Inf. Sci. & Technol. 43 1–43
[7] Han X P and Zhao J 2009 Web personal name disambiguation based on reference entity tables mined from the web Proc. of the 11th Int. Workshop on Web Inf. & Data Manag.-WIDM ’09 (New York: ACM) pp 75-82
[8] Cota R G, Ferreira A A, Nascimento C, Gonçalves M A and Laender A H F 2010 An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations J. of the Am. Soc. for Inf. Sci. & Technol. 61 1853-1870
[9] Kim S, Toutanova K and Yu H 2012 Multilingual named entity recognition using parallel data and metadata from Wikipedia Proc. of the 50th Annu. Meet. of the Assoc. for Comput. Linguist.: Long Papers (ACL ’12 vol 1) pp 694-702
[10] Lang J, Qin B, Song W, Liu L, Liu T and Li S 2009 Person name disambiguation of searching results using social network Chin. J. of Comput. 32 1365-1374
[11] Chiu J P C and Nichols E 2016 Named entity recognition with bidirectional LSTM-CNNs Trans. of the Assoc. for Comput. Linguist. 4 357-370
[12] Gridach M 2017 Character-level neural network for biomedical named entity recognition J. of Biomed. Inform. 70 85–91
[13] Qamas G K S, Yin J Z, Pan L M and Luo S L 2017 Research on the algorithm of named entity recognition based on deep neural network Netinfo Secur. 10 29-35
[14] Selvaperumal P and Suruliandi A 2016 Semi-supervised personal name disambiguation technique for the web Int. J. of Mod. Educ. & Comput. Sci. 8 28–36
[15] Zhang X, Chen F C and Huang R Y 2017 Research on entity disambiguation method based on fusion feature similarity Appl. Res. of Comput. 34 347-350
[16] Elkhidir M, Ibrahim M M, Khalid T A, Ibrahim S and Awadalla M 2015 Plagiarism detection using free-text fingerprint analysis 2015 World Symp. on Comput. Netw. and Inf. Secur. (WSCNIS) (Hammanet: IEEE) pp 1-4
[17] Ho P T and Sung K R 2014 Fingerprint-Based Near-Duplicate Document Detection with Applications to SNS Spam Detection Int. J. of Distrib. Sensor Netw. 10 40-44
[18] Manku G S, Jain A and Sarma A D 2007 Detecting near-duplicates for web crawling Proc. of the 16th Int. Conf. on World Wide Web (WWW ’07) (New York: ACM) pp 141-150
[19] Charikar M S 2002 Similarity estimation techniques from rounding algorithms Proc. of the Thirty-fourth Annu. Acm Symp. on Theory of Comput. (STOC ’02) (New York: ACM) pp 380-388
[20] https://www.biendata.com/competition/scholar2018/evaluation/