# Postprint: English Author Name Disambiguation Based on Sparse Distributed Representation

**Authors:** Zhai Xiaorui, Han Hongqi, Zhang Yunliang, Li Zhong

**Date:** 2018-10-11T00:00:00+00:00

## Abstract

To apply sparse distributed representation theory to author name disambiguation and understand its effectiveness in solving name disambiguation problems, we propose an author name disambiguation method for English literature based on sparse distributed representation. This method selects paper abstract text information as the disambiguation feature and generates it into binary SDR codes. Author name disambiguation is achieved by comparing the similarity between the SDR of the paper to be disambiguated and the SDRs of papers by authors with the same name. The final results are an accuracy of 98.21%, recall of 76.75%, and F-score of 86.17%, demonstrating that the proposed disambiguation method achieves favorable performance. Through comparison with methods that utilize co-author features for disambiguation, it is shown that this method can effectively resolve author name ambiguity issues in literature. Furthermore, this method can also identify papers whose authors are not included in the author database and assign them to new authors without requiring retraining or model updates.

## Full Text

## Preamble

### Research on English Author Name Disambiguation Based on Sparse Distributed Representation

*Zhai Xiaorui, Han Hongqi†, Zhang Yunliang, Li Zhong*

(Key Laboratory of Rich-media Knowledge Organization & Service of Digital Publishing Content, Institute of Scientific & Technical Information of China, Beijing 100038, China)

**Abstract:** To apply Sparse Distributed Representation (SDR) theory to author name disambiguation and evaluate its effectiveness in solving this problem,

this paper proposes a method for disambiguating English author names based on sparse distributed representation. The method selects paper abstracts as disambiguation features and generates binary SDR codes. Author name disambiguation is achieved by comparing the similarity between the SDR of a paper to be disambiguated and the SDRs of papers by authors with the same name. The final results show an accuracy of 98.21%, recall of 76.75%, and F-value of 86.17%, demonstrating that the proposed method achieves good performance. By comparing this method with a co-author feature-based approach, we show that our method can effectively solve the author name ambiguity problem. Additionally, this method can identify papers whose authors are not included in the author database and assign them to new authors without relearning or updating the model.

**Keywords:** name disambiguation; sparse distributed representation; semantic fingerprint; hierarchical temporal memory

## 0 Introduction

Name ambiguity is a serious problem in real-world applications where different individuals share identical names or the same person uses multiple name variations, significantly impacting effective information retrieval and utilization. As web data grows rapidly, manual resolution of name ambiguity has become impossible. Consequently, leveraging natural language processing and machine learning techniques to automatically eliminate name ambiguity has become a critical challenge that researchers must address, making it a prominent research topic in recent years. To advance automated name disambiguation technology, relevant institutions have organized specialized evaluation competitions, such as the Web People Search Evaluation Campaign (WePS) and CLP2010 (Chinese Language Processing 2010), which have generated significant impact and driven progress in this field.

Author name disambiguation in literature databases is a specific type of name ambiguity problem. When searching for a particular author' s publications in literature databases or institutional repositories, the system returns all works by authors sharing that name, reducing retrieval accuracy and causing confusion for users. Author name disambiguation aims to eliminate this ambiguity by determining which real-world entity each publication belongs to under a given name. This process is essential for constructing co-author social networks, evaluating research capabilities, and providing academic recommendations. Unlike web-based name disambiguation, author name disambiguation in literature requires richer personal information and comprehensive characteristics of individual researchers, going beyond simply identifying whether a webpage or article mentions a particular person.

Existing author name disambiguation methods primarily utilize short text information such as titles and keywords to represent authors' research directions or paper topics. However, they fail to leverage abstracts, which are long-text

information. Compared to short texts like keywords and titles, abstracts more completely express a paper' s core ideas, methodologies, and other crucial information, serving as a concentrated representation of the author's research themes. Although some clustering methods incorporate abstracts as a disambiguation feature, they typically convert long texts into word representations through tokenization, stop-word removal, and feature extraction, resulting in some loss of semantic information. Furthermore, most existing disambiguation methods rely on clustering algorithms for partitioning, but some clustering methods cannot determine the number of distinct individuals sharing a name, making it difficult to select the initial number of clusters—a choice that directly affects final clustering quality. Most current methods operate on existing datasets through classification or clustering, requiring relearning and model updates for newly added papers, which is inefficient and time-consuming.

Sparse Distributed Representation (SDR) theory is the primary information representation method in the Hierarchical Temporal Memory (HTM) theory of biological neural networks. It converts text information into binary sequences of fixed length (e.g., 16,384 bits) that carry semantic content features. In these sequences, "1" represents active bits, typically comprising 0.05% to 2% of the total bits. SDR transforms similarity comparisons between text contents into similarity measurements between binary sequence-based semantic fingerprints. Due to its high dimensionality and low sparsity, SDR exhibits high robustness and low false match rates, making it effective for matching different text contents. By leveraging these characteristics of SDR and the representativeness of abstracts for authors' research themes, we can identify a researcher' s body of work and thus achieve author name disambiguation. Therefore, this paper proposes an English author name disambiguation method based on SDR to explore the applicability and effectiveness of this theory in resolving name ambiguity.

## 1.1 Research Status of Name Disambiguation

Since Bagga and Baldwin first proposed cross-document name disambiguation in 1998, this problem has become a research hotspot among scholars worldwide. Through literature surveys in Wanfang Database and Web of Science, we categorize recent author name disambiguation methods into two main types: feature-based methods and machine learning-based methods.

Feature-based methods select personal information or bibliographic data as disambiguation features, converting paper similarity comparisons into feature comparisons to determine publication attribution. Commonly used personal information includes affiliation, email, and contact details, while bibliographic information includes titles, keywords, co-authors, and research directions. Singh extracted inventor surnames and regions from patent data, using if-else rules and string exact matching to determine whether two author record pairs matched correctly. Fleming et al. extracted and merged patentee and inventor region fields, using if-then-else matching rules and string exact matching with threshold settings to judge author record pairs. Xian Yantuan et al. selected proper

nouns, names, institutions, locations, titles, occupations, and concepts appearing as noun phrases in context as disambiguation features, constructing similarity matrices and employing affinity propagation clustering for disambiguation. Zhang Xiong et al. used co-author information, name correlation information, and topic information as features, implementing name disambiguation through multi-feature fusion. Li Mengya extracted three types of features from book author profiles: entity features, context features, and social relationship features, calculating author similarity using attribute mutual exclusion amplification and feature vacancy reduction methods, and completing disambiguation through agglomerative hierarchical clustering.

Machine learning-based methods can be further divided into three categories: supervised learning, unsupervised learning, and semi-supervised learning. Supervised learning methods train classifiers using labeled datasets for name disambiguation. Kim et al. employed random forest and DBSCAN clustering on USPTO (United States Patent and Trademark Office) inventor name disambiguation competition data, achieving results superior to the competition baseline. Unsupervised learning methods perform clustering in unlabeled training sets based on similarity calculations, treating pairs meeting threshold requirements as the same author. Zhu Liangliang utilized an improved K-means algorithm that selects initial cluster centers based on the max-min principle, overcoming the local convergence problem caused by random initial center selection in traditional K-means. Semi-supervised learning methods typically use small labeled datasets and large unlabeled datasets to train models for disambiguation. Ronald et al. employed Torvik and Smalheiser' s method to generate artificial labeled datasets with high accuracy through statistics and used logistic regression within a Bayesian framework to determine author record pair matching.

## 1.2 SDR Theory

Semantic fingerprint technology maps text content and features into fixed-length binary strings, typically 32-bit, 64-bit, or 128-bit, to represent text features, converting text similarity comparisons into distance measurements between binary sequences. This approach is commonly used in applications such as web deduplication and document plagiarism detection. Existing semantic fingerprint algorithms generate binary sequences of 32-bit, 64-bit, or 128-bit lengths, requiring minimal storage space but exhibiting low robustness.

SDR is a key component of the HTM theory proposed by Numenta and represents a form of semantic fingerprint. The SDR generation algorithm is based on semantic folding theory, which transforms text into semantic fingerprints. It obtains contextual text fragments of words from a constructed corpus and maps them onto a two-dimensional matrix, where text fragments with similar themes are positioned close together while those with different themes are far apart. By flattening this matrix into a one-dimensional vector, word-level SDR codes can be generated. Specifically, for a given word, if it appears in the corresponding text fragment, the corresponding position in the SDR vector is set to 1; other-

wise, it is 0. This vector represents the semantic meaning of the word based on its context, and individual word vectors can be combined to form sentence vectors and text vectors. SDR vectors are multi-dimensional, with length n typically ranging from 1,024 to 65,536, where the number of "1" bits ranges from 10 to 40, controlling sparsity between 0.05% and 2%. Each bit in an SDR carries certain semantic meaning; if two SDRs both have "1" at the same position, they share the attribute corresponding to that bit.

SDR storage requires only active bit information, significantly reducing storage space requirements. SDR theory defines similarity between two SDR codes using overlap—the count of positions where both vectors have "1". When overlap exceeds a threshold , the two SDRs are considered matching. When n=1,024 and =9, the false match probability between two SDR vectors drops to $3.0365\times10^{22}$, demonstrating SDR's high robustness and error tolerance. Even if some bits are discarded or moved, the represented semantics remain unchanged.

To facilitate SDR usage, Cortical.io provides the Retina API, which implements semantic fingerprint generation based on SDR theory. It converts input text information into a 128×128 dimensional matrix through semantic folding, representing it as a 16,384-bit SDR code. The API returns the indices of positions with value "1", which users can use to generate corresponding SDR codes.

## 2 English Author Name Disambiguation Method Based on SDR

The proposed English author name disambiguation method based on SDR determines whether a paper to be disambiguated belongs to a particular author by comparing its similarity with already-disambiguated papers, thereby associating the paper with the correct entity and distinguishing papers by authors with identical names. The method consists of five processing stages: SDR generation, SDR comparison, author matching, dispute arbitration, and author assignment, as shown in Figure 1 [Figure 1: see original paper].

### 2.1 Feature Selection and SDR Generation

Bibliographic information retrieved from literature databases typically includes title, author, co-authors, author affiliation, journal name, abstract, and keywords. Among these, the abstract is textual information that highly summarizes article content and represents the author' s ideas to a certain extent, making it an important feature for characterizing authors. Therefore, this study selects abstract information as the disambiguation feature. After obtaining the abstract, the SDR generation algorithm converts the abstract text into SDR codes as semantic fingerprints for disambiguation. The specific process is shown in Figure 2 [Figure 2: see original paper].

### 2.2 SDR Comparison

After generating the SDR for a paper's abstract, we compare it with the SDRs of all papers by existing authors with the same name. Similarity between two SDRs is calculated using the method provided by cortical.io, with the comparison result denoted as H(x). The value of H(x) is a decimal between 0 and 1. An H(x) value of 0 indicates that the two papers definitely belong to different authors, while a value of 1 indicates they definitely belong to the same author.

### 2.3 Matching Scheme

For a paper p to be disambiguated, we compare its SDR with the SDRs of N papers by a disambiguated author , obtaining N similarity comparison results Hi(x) (i=1, 2, ···, N). To determine whether p belongs to author , we set a similarity comparison threshold interval ( , ), where and are determined based on actual conditions. The process for determining whether paper p belongs to author is shown in Figure 3 [Figure 3: see original paper] and detailed as follows:

a) When the comparison result Hi(x) between paper p and the i-th paper of author exceeds threshold , we confirm that paper p' s author is . Although p undergoes N comparisons with author ' s N papers, this single occurrence is sufficient to assign paper p to author .

b) When all N SDR comparison results Hi(x) are less than threshold , two scenarios exist. First, if there exists Hi(x) such that < Hi(x) < , we consider that paper p may belong to author . Second, if all Hi(x) are less than , we consider that paper p cannot belong to author . We count the number of Hi(x) values falling within the interval ( , ), denoted as n. If n/N > h, then paper p is assigned to author , where h is a threshold parameter determined based on actual conditions.

### 2.4 Assignment Scheme

After matching, if paper p matches with m authors, three possible outcomes exist:

a) m = 0: Paper p fails to match any existing author and is assigned to a new author.

b) m = 1: Paper p matches only one author and is assigned to that author.

c) m > 1: Paper p matches multiple authors simultaneously. In this case, an arbitration procedure determines which author should receive the paper.

Without loss of generality, assume paper p matches both author and author . We calculate the average of similarity comparison results between paper p

and all papers by each author:

$$\frac{\sum H(\alpha_1)}{N_{\alpha_1}} \text{ and } \frac{\sum H(\alpha_2)}{N_{\alpha_2}}$$

If

$$\frac{\sum H(\alpha_1)}{N_{\alpha_1}} > \frac{\sum H(\alpha_2)}{N_{\alpha_2}}$$

the paper is assigned to author ; otherwise, it is assigned to author (see Figure 4 [Figure 4: see original paper]). For cases with more than two matching authors, the author with the highest average similarity comparison result is selected as the assignment target.

### 2.5 Evaluation Metrics

To evaluate the effectiveness of the SDR-based name disambiguation method, we adopt the precision, recall, and F-measure metrics used in the Chinese Language Processing International Conference (CLP-2012) hosted by the Chinese Information Processing Society of China (CIPS) and the ACL Special Interest Group for Chinese Language Processing (SIGHAN) in 2012.

Precision refers to the ratio of papers correctly identified as belonging to author among all papers assigned to , while recall refers to the ratio of papers correctly identified as belonging to author among all papers that actually belong to . Both values range from 0 to 1, with values closer to 1 indicating better performance. Since these two metrics often influence each other in practice—improving one typically reduces the other—we use the F-value to comprehensively reflect overall performance.

If we view name disambiguation results as clusters where each cluster represents the same author' s paper set, the precision and recall for each cluster are calculated as follows:

$$\text{Precision}_i = \frac{|R_i \cap S_i|}{|S_i|}$$

$$\text{Recall}_i = \frac{|R_i \cap S_i|}{|R_i|}$$

where R represents the manually disambiguated result set, R R represents a particular cluster in the manual results, S represents the method-disambiguated result set, and S S represents a particular cluster in the method results. The sizes of the two sets are denoted as |R | and |S |, respectively.

After obtaining precision and recall for each cluster, we calculate their averages as the overall precision and recall, where N represents the number of clusters:

$$\text{Precision} = \frac{1}{N}\sum_{i=1}^{N}\text{Precision}_i$$

$$\text{Recall} = \frac{1}{N}\sum_{i=1}^{N}\text{Recall}_i$$

The overall F-value is then calculated as:

$$F\text{-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

## 3.1 Dataset Construction

To validate the proposed method, we constructed an experimental dataset manually using the ResearchGate academic social network. We selected four highly ambiguous names: J Huang, L Stevens, T Joe, and J Baker (one Chinese author' s English name and three foreign names). After preliminary analysis of the collected author paper information, we found that some literature data lacked abstracts, so these were removed from the dataset. The final dataset contains 88 papers by 19 authors. The dataset was divided into two parts: Dataset 1 and Dataset 2. Dataset 1 includes 47 papers by 17 authors, while Dataset 2 includes 41 papers by 18 authors. We then extracted portions from both datasets to form training sets, labeled as Training Set 1 (D1) and Training Set 2 (D2). D1 contains 23 papers by 7 authors, and D2 contains 19 papers by 7 authors. D1 and D2 are used to estimate threshold parameters , , and h, while Dataset 1 and Dataset 2 are used for testing the proposed SDR-based name disambiguation method.

For experimental convenience, we used Cortical.io's Retina API to generate SDR representations for each paper' s abstract. The Retina API does not require text tokenization or stop-word removal. However, special characters in actual literature or non-English symbols introduced by authors from non-English-speaking countries can affect SDR results, necessitating appropriate normalization of abstract text. Figure 5 [Figure 5: see original paper] shows the numeric sequence returned by the Retina API, where each number represents the index of a position with value "1" in the SDR vector. Figure 6 [Figure 6: see original paper] shows the actual SDR code generated from this sequence.

By comparing the SDR codes of papers by authors with the same name in D1 and D2, we obtained a similarity result matrix. We analyzed similarity results between any two papers by the same author and between any two papers by different authors, as shown in Figure 7 [Figure 7: see original paper]. The results within the bold boxes represent similarities between papers by the same author, while the remaining values represent similarities between papers by different authors.

Analysis reveals that similarity results between any two papers by the same author fall within the range (0.1522, 0.5833), primarily concentrated in (0.42, 0.52). Similarity results between any two papers by different authors fall within (0.1657, 0.4919), primarily concentrated in (0.29, 0.44), as shown in Figure 8 [Figure 8: see original paper]. Therefore, we set the threshold selection interval to (0.42, 0.52), with = 0.42.

To determine the optimal parameters and h, we further analyzed the similarity matrix. We varied from 0.42 to 0.52 in increments of 0.01 while setting h to 20%, 30%, and 40%, observing the resulting precision, recall, and F-value curves to find the best parameter combination. As shown in Figure 9 [Figure 9: see original paper], the optimal threshold combination corresponds to = 0.42, = 0.50, and h = 20%.

## 3.3 Experimental Results of SDR-Based Disambiguation

Based on the experimental protocol, we compared the SDR codes of any two papers by authors with the same name in Dataset 1 and Dataset 2 to obtain the similarity matrix. Using the derived thresholds ( = 0.42, = 0.50, h = 20%), we performed name disambiguation. The assignment rules were: if H(x) > 0.50, assign the paper to that author; if 0.42 < H(x) < 0.50, calculate the percentage of H(x) values within (0.42, 0.50); if this percentage exceeds 20%, match with the corresponding author. If a paper matches only one author, assign it to that author. If it matches multiple authors, compare the average H(x) values and assign the paper to the author with the higher average. If no match is found with existing authors, assign the paper to a new author.

The final experimental results show an accuracy of 98.21%, recall of 76.75%, and F-value of 86.17%.

## 3.4 Comparison with Co-author Feature-Based Disambiguation

In bibliographic data, besides the abstract text used in this study, other features are widely used for name disambiguation, such as co-author and affiliation features. According to Zhang Xiong et al.' s research, co-author feature-based disambiguation achieves good results. Therefore, we conducted experiments using co-author features for comparison to evaluate the effectiveness of our SDR-based approach.

In the co-author feature-based name disambiguation experiment, we first manually normalized co-author names to avoid ambiguity issues affecting results, then used string matching for disambiguation. If two papers by authors with the same name share at least one co-author, we considered them to be by the same person.

The experiment yielded an accuracy of 98.32%, recall of 73.68%, and F-value of 84.24%. Figure 10 [Figure 10: see original paper] compares the two methods.

The accuracy difference is minimal, with the co-author method slightly higher, but our SDR-based method shows a clear advantage in recall, resulting in a better F-value. The dataset used in this experiment contains few single-author papers, which contributes to the co-author method' s good accuracy. With a larger proportion of single-author papers, the co-author method' s accuracy would likely decrease significantly.

For papers whose authors are not included in Dataset 1, the co-author feature method cannot perform disambiguation, whereas our proposed method can identify such papers and assign them to new authors. However, both methods fail for cases where authors with the same name have co-author relationships.

## 4 Conclusion

Author name disambiguation is fundamental for research output evaluation, co-author social network construction, and knowledge service system development. This paper proposes an English author name disambiguation method based on sparse distributed representation. The method selects abstracts as disambiguation features and uses the SDR generation algorithm to produce 16,384-bit SDR codes. Through similarity comparisons of SDR codes and threshold parameter determination, papers meeting the criteria are assigned to corresponding authors. The final experimental results demonstrate an accuracy of 98.21%, recall of 76.75%, and F-value of 86.17%, proving that sparse distributed representation can be effectively applied to name disambiguation. The proposed method can successfully identify papers whose authors are not in the author database and assign them to new authors.

Although this study achieved good results on the constructed dataset, preliminary validating the feasibility and effectiveness of SDR-based author name disambiguation, several limitations remain. First, the experimental dataset is small-scale and may not cover various real-world scenarios of author name ambiguity, potentially lacking comprehensiveness and representativeness. Second, while the method assigns some authors to new author categories, it does not synchronize these newly discovered authors back to the disambiguated database in real time. Although this delayed update did not significantly impact results in our experiments, it could affect disambiguation performance in large-scale scenarios where some authors assigned to new categories might actually be the same person.

This method is an improved version of Fu Yuan' s approach, with the following main differences: a) Different semantic fingerprint generation methods. Fu Yuan' s method uses hash functions to generate hash values for words in paper texts and produces semantic fingerprints through the Simhash algorithm. Our method generates word SDRs based on large-scale corpus learning, and paper text SDR fingerprints are generated based on word SDRs. b) Different fingerprint matching schemes. Fu Yuan' s experiments used one threshold and parameter h, identifying authors as the same when the proportion of fingerprint

comparison results exceeding  surpassed h. Our method uses three threshold parameters ( ,  , and h): if comparison results exceed  , they are considered the same author; if results fall between  and  , the proportion exceeding h determines authorship. c) Different language environments: Fu Yuan' s work disambiguated Chinese author names, while this study focuses on English author names.

Future research will focus on: a) Integrating co-author features, affiliation features, and SDR for improved disambiguation performance to facilitate practical system application; b) Applying deep learning algorithms for threshold selection optimization to achieve better disambiguation results; c) Currently, the core code for SDR generation is not publicly available, so we can only obtain English text SDR representations through Cortical.io' s Retina API. We will construct a Chinese corpus and research Chinese word SDR generation based on Chinese corpora to enable Chinese literature author name disambiguation.

## References

[1] Wang Xin. Research and implementation of key techniques of personal name disambiguation [D]. Harbin: Harbin Institute of Technology, 2012.

[2] Fu Yuan, Zhu Lijun, Han Hongqi. A survey of name disambiguation [J]. Technology Intelligence Engineering, 2016 (1): 53-58.

[3] Piotr A, Szymon S. Person name disambiguation for building university knowledge base [C]// Proc of Asian Conference on Intelligent Information and Database Systems. Berlin: Springer, 2016: 270-279.

[4] Ren Jinghua. Document author name disambiguation by using optimized DBSCAN algorithm [J]. Library Theory and Practice, 2014 (12): 61-65.

[5] Yuan Junpeng, Yu Zhenglu, Su Cheng, et al. A survey of author name disambiguation [J]. Library Theory and Practice, 2011 (10): 60-65.

[6] Neil R S, Vetle I T. Author name disambiguation [J]. Annual Review of Information Science & Technology, 2015, 43 (1): 1-43.

[7] Zhang Xiong, Chen Fucai, Huang Ruiyang. Research on entity disambiguation method based on fusion feature similarity [J]. Application Research of Computers, 2017, 34 (2): 347-350.

[8] Zhu Liangliang. Research on name disambiguation based on improved K-means algorithm [J]. Software Guide, 2013 (5): 63-66.

[9] Du Jingjun. Research on method of Chinese named entity disambiguation based on Chinese Wikipedia [D]. Hangzhou: Hangzhou Dianzi University, 2012.

[10] Yang Yilin, Chen Gang, Zhou Jie, et al. Research on name disambiguation: A survey [J]. Journal of Information Engineering University, 2016 (4): 478-483.

[11] Chen Chen, Wang Houfeng. Social network based cross-document personal name disambiguation [J]. Journal of Chinese Information Processing, 2011, 25 (5): 75-82.

[12] Yang Yilin, Zhou Jie, Li Bicheng. Name disambiguation algorithm based on ensemble [J]. Application Research of Computers, 2016, 33 (9): 2716-2720.

[13] Ahmad S, Hawkins J. Properties of sparse distributed representations and their application to hierarchical temporal memory [J]. Eprint Arxiv, 2015.

[14] De Sousa-Webber F. Semantic folding theory and its application in semantic fingerprinting [J]. Computer Science, 2015.

[15] Amit B, Breck B. Entity-based cross-document coreferencing using the Vector Space Model [C]// Proc of Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 1998.

[16] Jasjit S. Collaborative networks as determinants of knowledge diffusion patterns [J]. Management Science, 2005, 51 (5): 756-770.

[17] King III C, Fleming L, Juda A. Small worlds and regional innovative advantage [J]. Organization Science, 2007 (6): 938-954.

[18] Xian Yantuan, Yu Zhengtao, Hong Xudong, et al. Collaborative entity disambiguation method based on weighted feature overlap relatedness for Chinese [J]. Journal of Chinese Information Processing, 2017, 31 (2): 36-41.

[19] Li Mengya. Research on Chinese book author' s name disambiguation based on fusion features [J]. Computer Knowledge and Technology, 2018 (11): 182-184.

[20] Kim K, Khabsa M, Giles C L. Random forest DBSCAN for USPTO inventor name disambiguation [EB/OL]. (2017-09-14). https://arxiv.org/abs/1602.

[21] Li Guancheng, Lai R, D' Amour A, et al. Disambiguation and co-authorship networks of the U.S. patent inventor database (1975-2010) [J]. Research Policy, 2014, 43 (6): 941-955.

[22] Torvik V I, Smalheiser N R. Author name disambiguation in MEDLINE [J]. ACM Trans on Knowledge Discovery from Data, 2009, 3 (3): 1-29.

[23] Fu Yuan. Research on Chinese authors name disambiguation based on semantic fingerprint [D]. Beijing: Institute of Scientific and Technical Information of China, 2016.

[24] Cui Yuwei, Ahmad S, Hawkins J. The HTM spatial pooler-A neocortical algorithm for online sparse distributed coding [J]. Frontiers in Computational Neuroscience, 2017, 11: 111.

[25] Hawkins J, Ahmad S, Purdy S, et al. Biological and machine intelligence (BAMI) [EB/OL]. (2018-03-08) [2018-06-27]. https://numenta.com/resources/biological-and-machine-intelligence/.

[26] Yin Xiangquan, Zeng Shan, Mi Kai. Personal name disambiguation-based research on self-citation statistics [J]. Information Research, 2015 (5): 57-59, 67.

[27] Cortical.io. retina-sdk.py: A python client for the cortical.io retina API [EB/OL]. (2017-03-06) [2018-06-27]. https://github.com/cortical-io/retina-sdk.py.

*Note: Figure translations are in progress. See original paper for figures.*