Strong Baselines for Author Name Disambiguation with and without Neural Networks

Zhenyu Zhang, Bowen Yu, Tingwen Liu, Dong Wang

Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

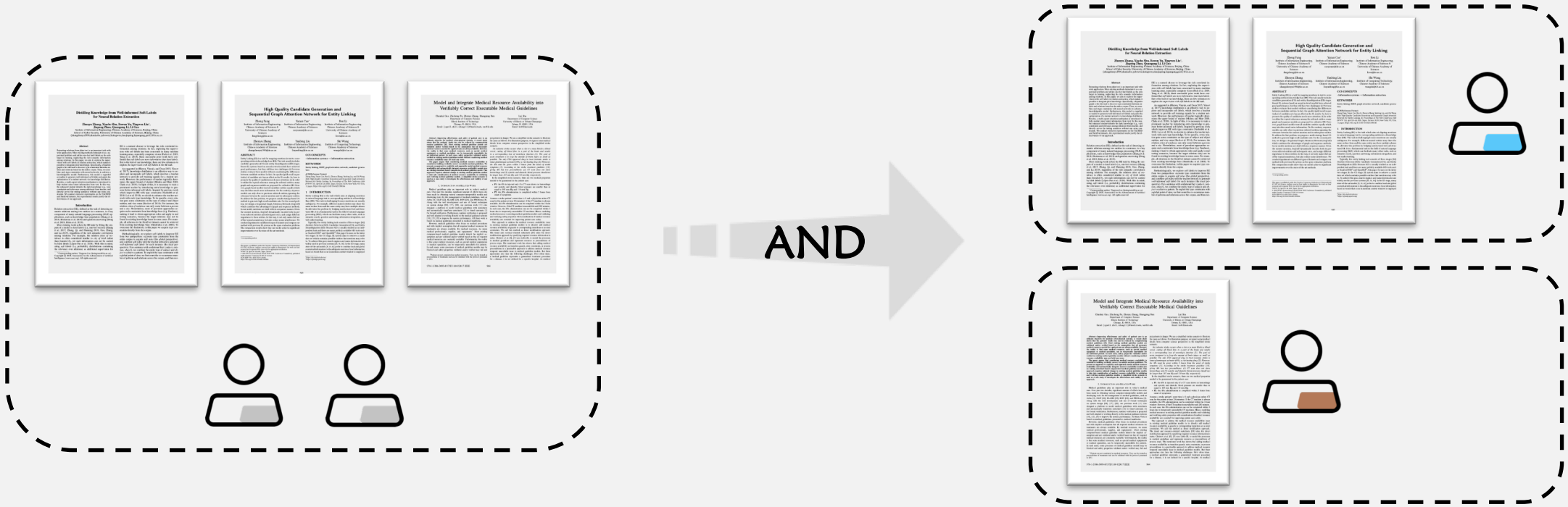
PAKDD 2020, Singapore, May 2020

OUTLINE

- Background
- Our Work
- Experiments
- Conclusion

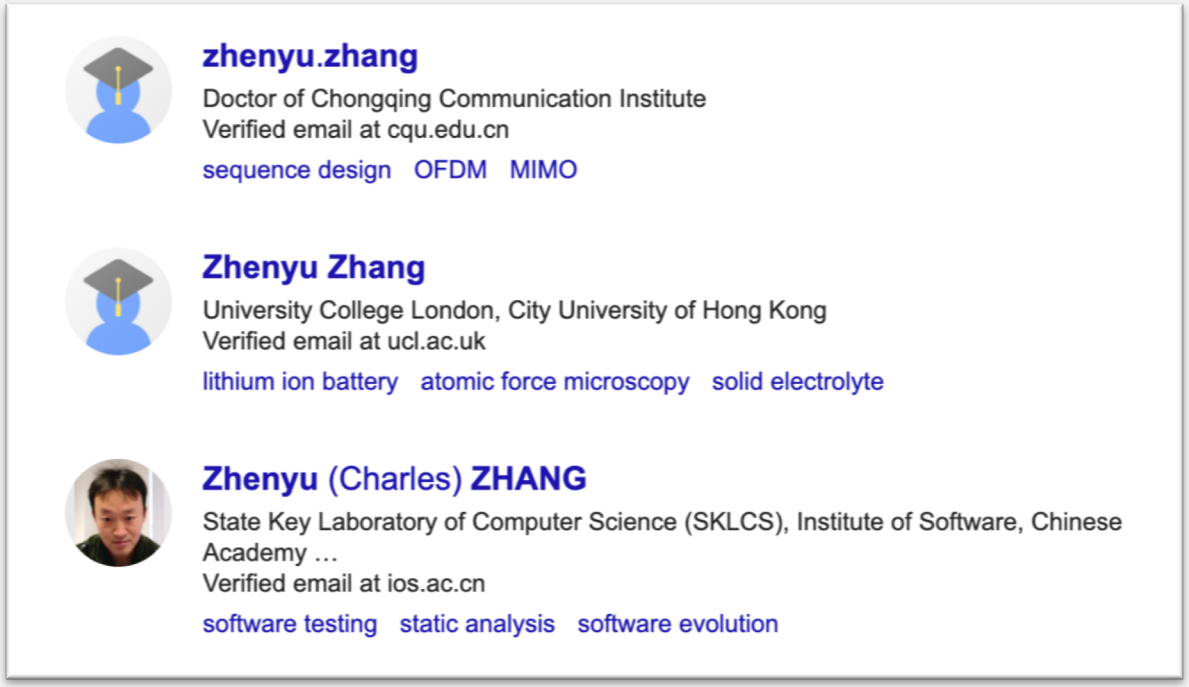
Author Name Disambiguation

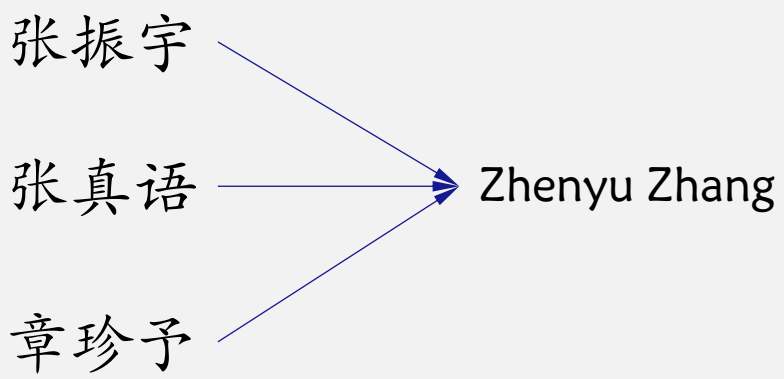
Author Name Disambiguation (AND), is the task of clustering unique authors using the metadata of publication records (such as title, venue, author name and affiliation). The task of disambiguating "***who is who***" in scientometrics is beneficial to information retrieval, bibliographic data analysis and other follow-up work.



Main Challenge

AND is not an elementary problem because different authors may share the same name, which is quite common for Asians, especially different Chinese names will be the same when mapped to English.





Existing Solutions

AND is typically *viewed as a clustering problem* and solved by clustering models, such models have to answer two questions inevitably:

- **How to quantify the similarity?**
  (1) Feature-based methods.
  (2) Graph-based methods.
  (3) Neural-based methods.
  Many of them involve complex and time-consuming architectures that yield smaller gains over previous state-of-the-art.
- **How to determine the cluster size?**
  (1) Assume the number of clusters is known beforehand.
  (2) Predict the number by applying another model.
  The former is unrealistic in real situations and the latter may lead to error propagation.

Motivation

AND IS ***NOT*** A TYPICAL CLUSTERING TASK

- **We should pay more attention to the precision.**
  Once two clusters are merged incorrectly, re-splitting them is an almost impossible process. Thus, cost-effective blocking technique and lightweight rule-based methods are worthy of research as they have been proven to achieve convincing precision.

- **Some complex architectures are inconsequential.**
  We peel away unnecessary components until arrive at the simplest model that works well without any prior knowledge about cluster size.

OUTLINE

- Background
- Our Work
- Experiments
- Conclusion

Problem Definition

Given an author name α and a set of publication records $\mathcal{P} = \{p_1, p_2, ..., p_l\}$ with the name α, the problem of author name disambiguation is to partition the publication records $\mathcal{P}$ into different clusters $\{C_1, C_2, ..., C_K\}$ such that:

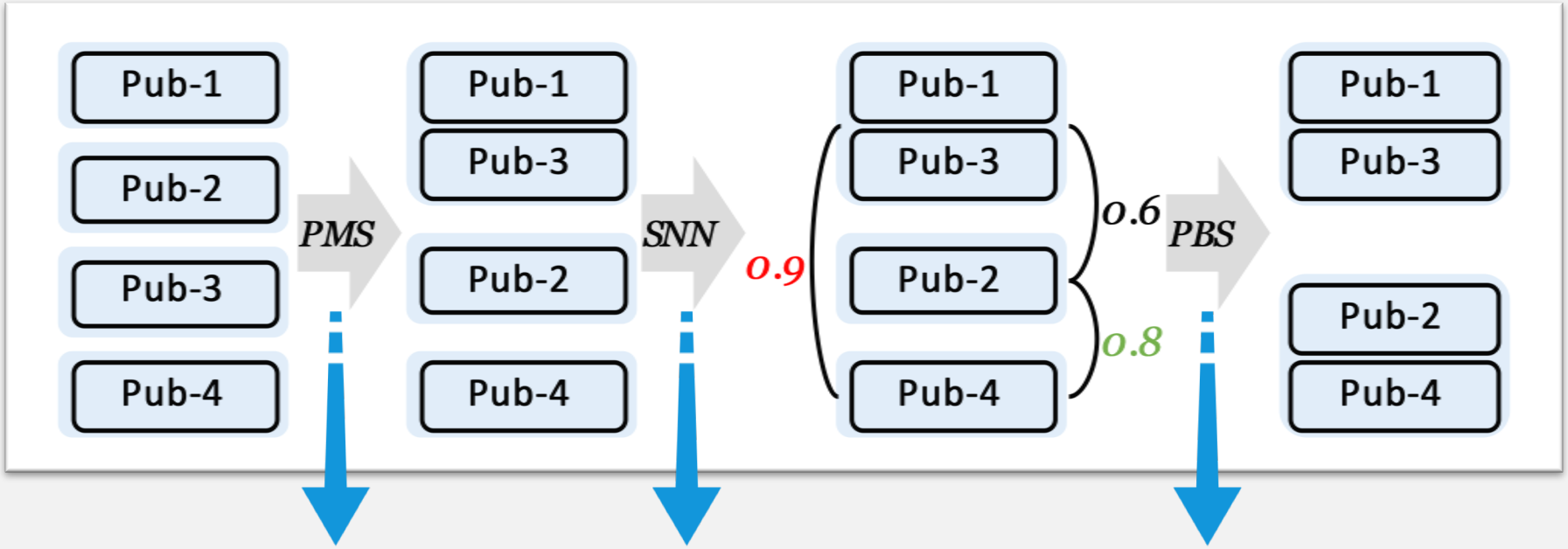
- All the records in C_k belong to the same author α_k.
- All the records in $\mathcal{P}$ by α_k are in C_k.

where $\{\alpha_1, ..., \alpha_K\}$ are K different people with the same name α.

With this in mind, we first scatter the publication records $\mathcal{P} = \{p_1, p_2, ..., p_l\}$ into l sets, and there is only one unique publication p in each original set.

Our Model

Our design philosophy is based on the observation that researchers have stable coauthors, and publication topics belong to one researcher should be close in the semantic space in a certain period.



Pre-Merging Strategy

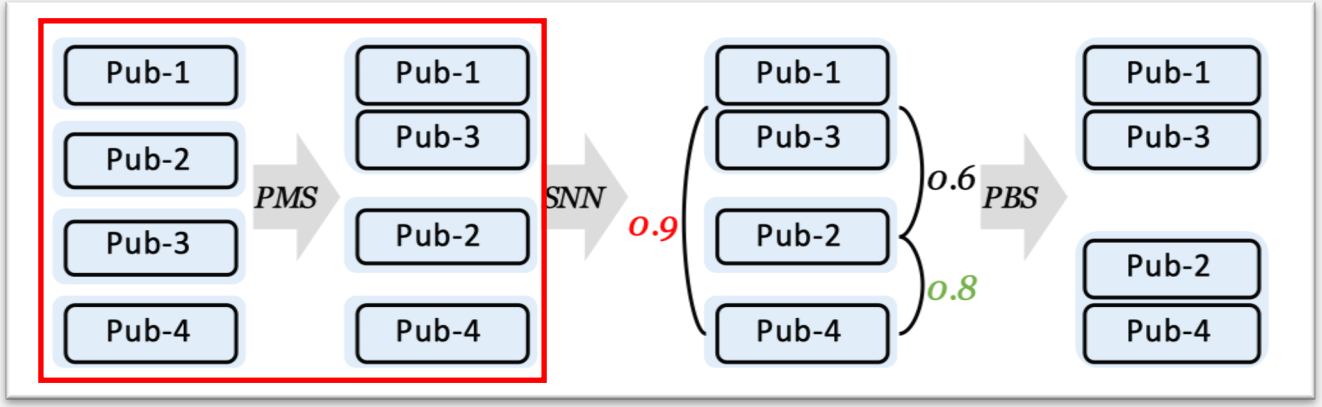
This step aims to merge the initial publication sets preliminarily using point-to-point and cluster-to-cluster rules.

➢ point-to-point

Given two publications p_i and p_j, if $|S_n(p_i) \cap S_n(p_j)| > \lambda_1$, or $A_\alpha(p_i) = A_\alpha(p_j)$ & $|S_n(p_i) \cap S_n(p_j)| > 1$, then $\mathcal{M}(p_i, p_j) = 1$. For a publication p_i, $S_n(p_i)$ and $A_\alpha(p_i)$ denote the set of author names and the affiliations of author name α.

(1) we regard two publications p_i, p_j belong to the same cluster when the coauthor number of p_i and p_j **exceeds a threshold λ_1**.

(2) if affiliations of name α are identical in p_i and p_j, the **threshold is relaxed to 1**.



Pre-Merging Strategy

➢ cluster-to-cluster

Given two clusters C_i and C_j, if $\mathcal{O}_n(C_i, C_j) > \lambda_2$, or $\mathcal{O}_a(C_i, C_j) > \lambda_2$, then $\mathcal{M}(C_i, C_j) = 1$, where $\mathcal{O}_x(C_i, C_j)$ denotes the overlap ratio of two clusters in the aspect of x, and $x \in \{n, a\}$ denotes the name or affiliation of authors. We define the overlap ratio $\mathcal{O}_x(C_i, C_j)$ as:

$$\mathcal{O}_x(C_i, C_j) = \frac{\sum_{\bar{x} \in (S_x(C_i) \cap S_x(C_j))} (F_{\bar{x}}(C_i) + F_{\bar{x}}(C_j))}{\min(\sum_{\bar{x} \in S_x(C_i)} F_{\bar{x}}(C_i), \sum_{\bar{x} \in S_x(C_j)} F_{\bar{x}}(C_j))}$$

where $F_{\bar{x}}(C_i)$ is the occurrence number of $\bar{x}$ in the cluster C_i.

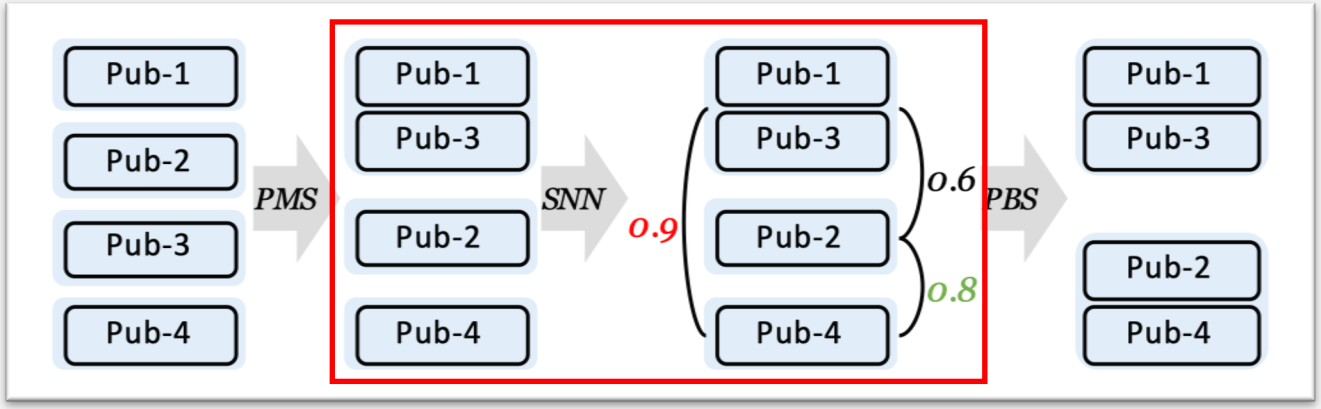
Here we take author name as an example to explain the overlap ratio.

(1) The **numerator** is the total occurrence number of overlap author names in author name sets $S_n(C_i)$ and $S_n(C_j)$.

(2) The **denominator** is the minimum of total occurrence number of all author names in $S_n(C_i)$ and $S_n(C_j)$.

Simple Neural Networks

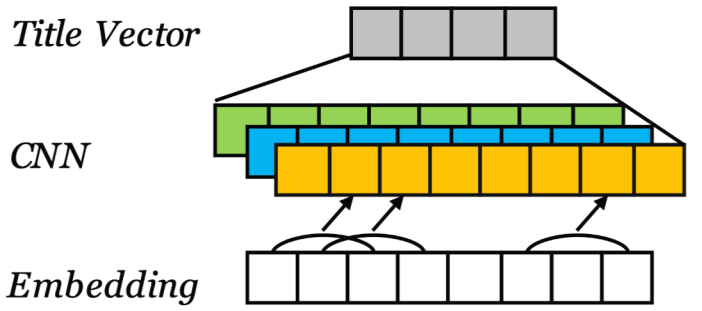
This step aims to determine whether two publications belong to the same author based on the **topic similarity**.



(1) Transform the title of publication into a word sequence.
(2) Utilize a standard CNN to produce title vector for each publication.
(3) Train the representation model $\mathcal{R}(\cdot)$ with a triplet loss.
(4) Calculate cluster embeddings with mean pooling.

$$\mathcal{L} = \frac{1}{|\mathcal{T}|}\sum_{i=1}^{|\mathcal{T}|}\max(0, \delta - \cos(\mathcal{R}(p_i), \mathcal{R}(p_{i_+})) + \cos(\mathcal{R}(p_i), \mathcal{R}(p_{i_-})))$$

$$\mathbf{c}_i = \frac{1}{|C_i|}\sum_{j=1}^{|C_i|}\mathcal{R}(p_j)$$



Isomorphic Transfer of Syntactic Structures in Cross-Lingual NLP

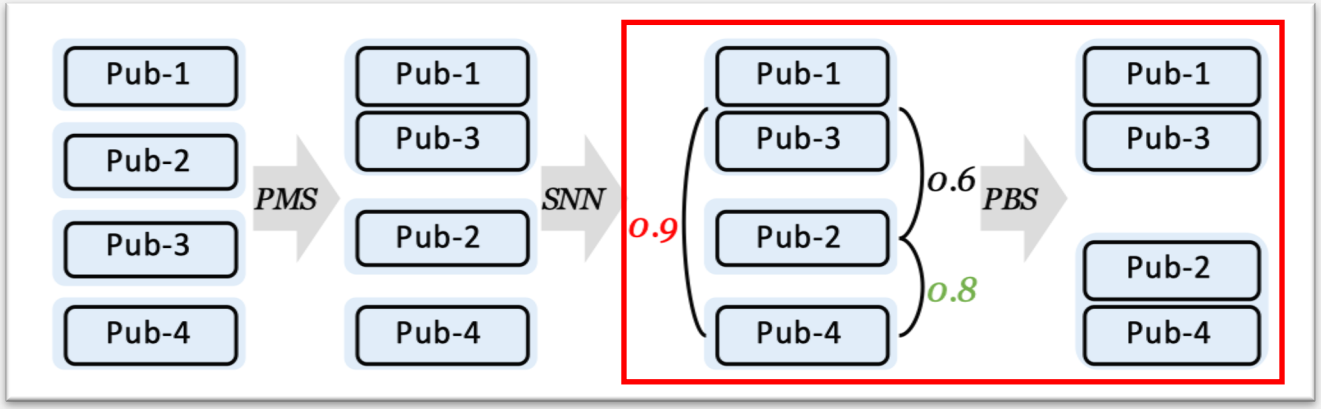
Post-Blocking Strategy

Based on the learned representation model and cluster embeddings, this step is proposed to make the final partition.

➢ anchor-to-anchor

if $S_n(p_i^*) \cap S_n(p_j^*) = \{\alpha\}$ and $S_{a\setminus\alpha}(p_i^*) \cap S_{a\setminus\alpha}(p_j^*) = \emptyset$, then $\mathcal{M}(C_i, C_j) = 0$.
For an anchor publication p_i^*, $S_{a\setminus\alpha}(p_i^*)$ denote the set of affiliations except the current author α.

If there is no intersection between the name sets or affiliation sets of p_1^* and p_2^* except name α and its affiliation, these two clusters do not belong to the same author.



OUTLINE

- Background
- Our Work
- Experiments
- Conclusion

Overall Performance

- **Aminer** (500/100 author names for training/testing)

Table 1. Results of author name disambiguation on the AMiner benchmark dataset. † marks results reported in [25].

Model	Precision	Recall	F1 Score
Basic Rules [25]†	44.94	89.30	53.42
Fan et al. [3]†	81.62	40.43	50.23
Louppe et al. [13]†	57.09	**77.22**	63.10
Zhang and Al Hasan [24]†	70.63	59.53	62.81
Zhang et al. [25]†	77.96	63.03	67.79
PMS	**81.86**	55.61	66.23
PMS+SNN	73.90	61.97	67.41
PMS+SNN+PBS (PNP)	76.92	64.54	**70.19**

We use an optimal reject threshold (0.8) tuned on training set.

[25] Zhang et al. *Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop*. SIGKDD. 2018.

Efficiency Analysis

Runtime & Parameter Number

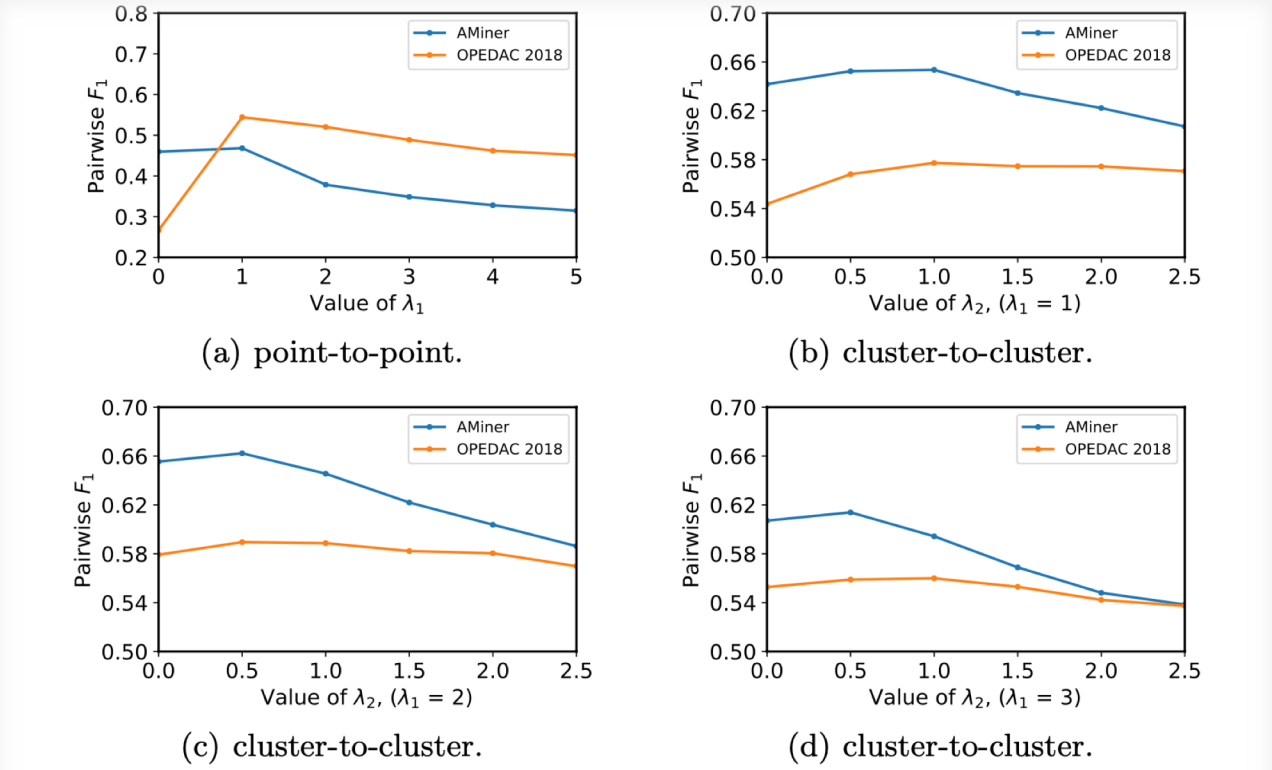
Table 2. Runtime and trainable parameter number of different models.

Model	Runtime		Trainable
	Training	Testing	Parameters Number
Zhang el al. [25]	>24h	~573s	3,024,193
PMS	-	~31s	0
SNN	~2h	-	30,100
PBS	-	~88s	0
PMS+SNN+PBS (PNP)	~2h	~119s	30,100

[25] Zhang et al. *Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop*. SIGKDD. 2018.

Rule Sensitivity Analysis

- **Open Academic Data Challenge 2018** (50 author names for testing)



(a) point-to-point. (b) cluster-to-cluster.

(c) cluster-to-cluster. (d) cluster-to-cluster.

Fig. 3. The effects of λ_1 and λ_2 to the performance when using different datasets. The subfigure (a) depicts results when only the point-to-point rules are used, while remaining subgraphs show the results of the entire pre-merging strategy.

OUTLINE

- Background
- Our Work
- Experiments
- Conclusion

Conclusion

- We take heuristic rules that come from real-world observation into consideration and propose strong baselines for AND task without any knowledge about cluster size.
- Experimental results verify the advantage of our approach over state-of-the-art models, and heuristic rules can be extended to other real-world datasets.
- SUGGESTION: The most important thing is to consider baselines that do not involve complex architectures, simple methods might lead to unexpected performances.

Thanks!

Questions and Advices?