



# Revisiting Name Disambiguation

## - Focus on unassigned papers

Jing Zhang

Information School, Renmin University of China

Collaborate with Bo Chen (*RUC*), Jie Tang (*THU*)

# Google Scholar

Google 学术搜索

## Zhigang Wang

Tsinghua University
在 mails.tsinghua.edu.cn 的电子邮件经过验证

Knowledge Graph

关注

One common coauthor “Jie Tang”

| 标题 | 引用次数 | 年份 |
| --- | --- | --- |
| Paclitaxel-loaded and A10-3.2 aptamer-targeted poly (lactide-co-glycolic acid) nanobubbles for ultrasound imaging and therapy of prostate cancer<br>M Wu, Y Wang, Y Wang, M Zhang, Y Luo, J Tang, Z Wang, D Wang, ...<br>International journal of nanomedicine 12, 5313 | 9 | 2017 |
| Domain specific cross-lingual knowledge linking based on similarity flooding<br>L Pan, Z Wang, J Li, J Tang<br>International Conference on Knowledge Science, Engineering and Management … | 1 | 2016 |
| Boosting to Build a Large-Scale Cross-Lingual Ontology<br>Z Wang, L Pan, J Li, S Li, M Li, J Tang<br>China Conference on Knowledge Graph and Semantic Computing, 41-53 | | 2016 |

# DBLP

## Jing Zhang

**2009**

- [c14] Jie Tang, Jing Zhang, Jeffrey Xu Yu, Zi Yang, Keke Cai, Rui Ma, Li Zhang, Zhong Su: **Topic Distributions over Links on Web.** ICDM 2009: 1010-1015
- [c13] Jie Tang, Jing Zhang: **A Discriminative Approach to Topic-Based Citation Recommendation.** PAKDD 2009: 572-579
- [c12] Zi Yang, Jie Tang, Jing Zhang, Juanzi Li, Bo Gao: **Topic-Level Random Walk through Probabilistic Model.** APWeb/WAIM 2009: 162-173

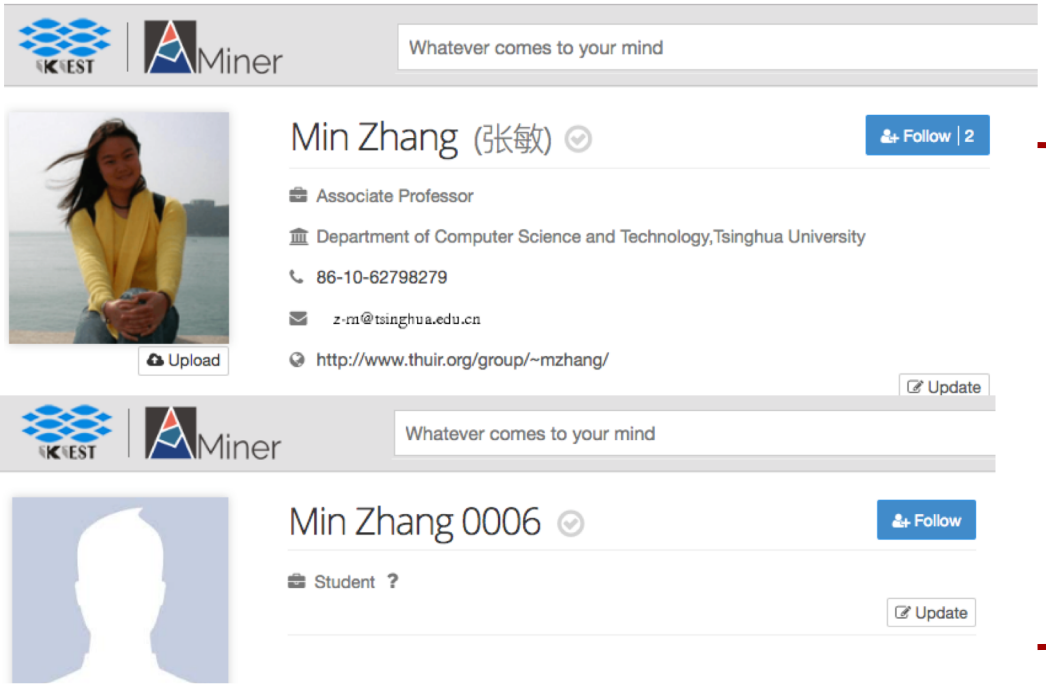
**2008**

- [j1] Juanzi Li, Jie Tang, Jing Zhang, Qiong Luo, Yunhao Liu, MingCai Hong: **Arnetminer: expertise oriented search using social networks.** Frontiers Comput. Sci. China 2(1): 94-105 (2008)
- [c11] Jie Tang, Ruoming Jin, Jing Zhang: **A Topic Modeling Approach and Its Integration into the Random Walk Framework for Academic Search.** ICDM 2008: 1055-1060
- [c10] Dongming Li, Jing Zhang, Li Zhang, Yaju Liu, Bo Du: **An Automatic Voice Query System for Bank Based on Telephone Network.** ISIP 2008: 629-632
- [c9] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, Zhong Su: **ArnetMiner: extraction and mining of academic social netwo**
- [c8] Jing Zhang, Jie Tang, Liu Liu, Juan-Zi Li: **A Mixture Model for Expert Finding.** PAKDD 2008: 466-478
- [c7] Jing Zhang, Jie Tang, Bangyong Liang, Zi Yang, Sijie Wang, Jingjing Zuo, Juanzi Li: **Recommendation over a Heterogeneous Social Network.** WAIM 2008: 309-316
- [c6] Feng Wang, Juanzi Li, Jie Tang, Jing Zhang, Kehong Wang: **Name Disambiguation Using Atomic Clusters.** WAIM 2008: 357-364
- [c5] Jie Tang, Jing Zhang, Limin Yao, Juan-Zi Li:

One common coauthor “Li Zhang”

# AMiner

## Min Zhang




Papers are over-partitioned

9

**Explicit factor models for explainable recommendation based on phrase-level sentiment analysis**

Yongfeng Zhang, Guokun Lai, **Min Zhang 0006**, Yi Zhang, Yiqun Liu, Shaoping Ma

SIGIR (2014)

**Cited by 216** Bibtex

8

**Improving Tail Query Performance by Fusion Model**

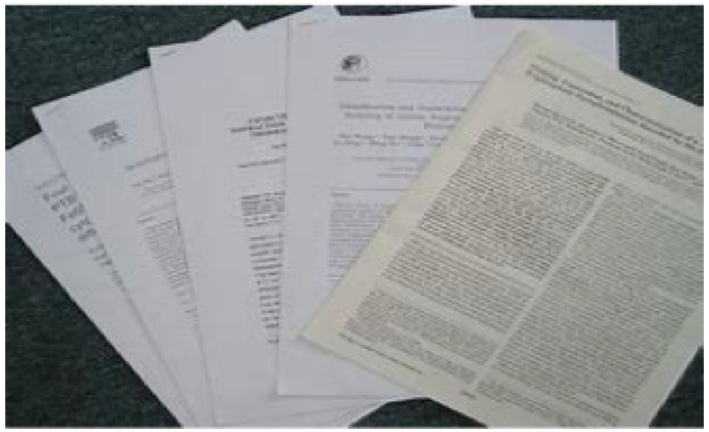
Shuai Huo, **Min Zhang 0006**, Yiqun Liu, Shaoping Ma

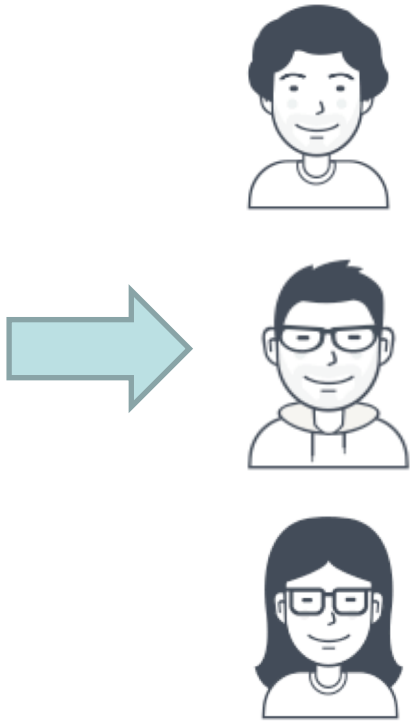
CIKM (2014)

**Cited by 3** Bibtex

# Beyond “Name Disambiguation From Scratch”

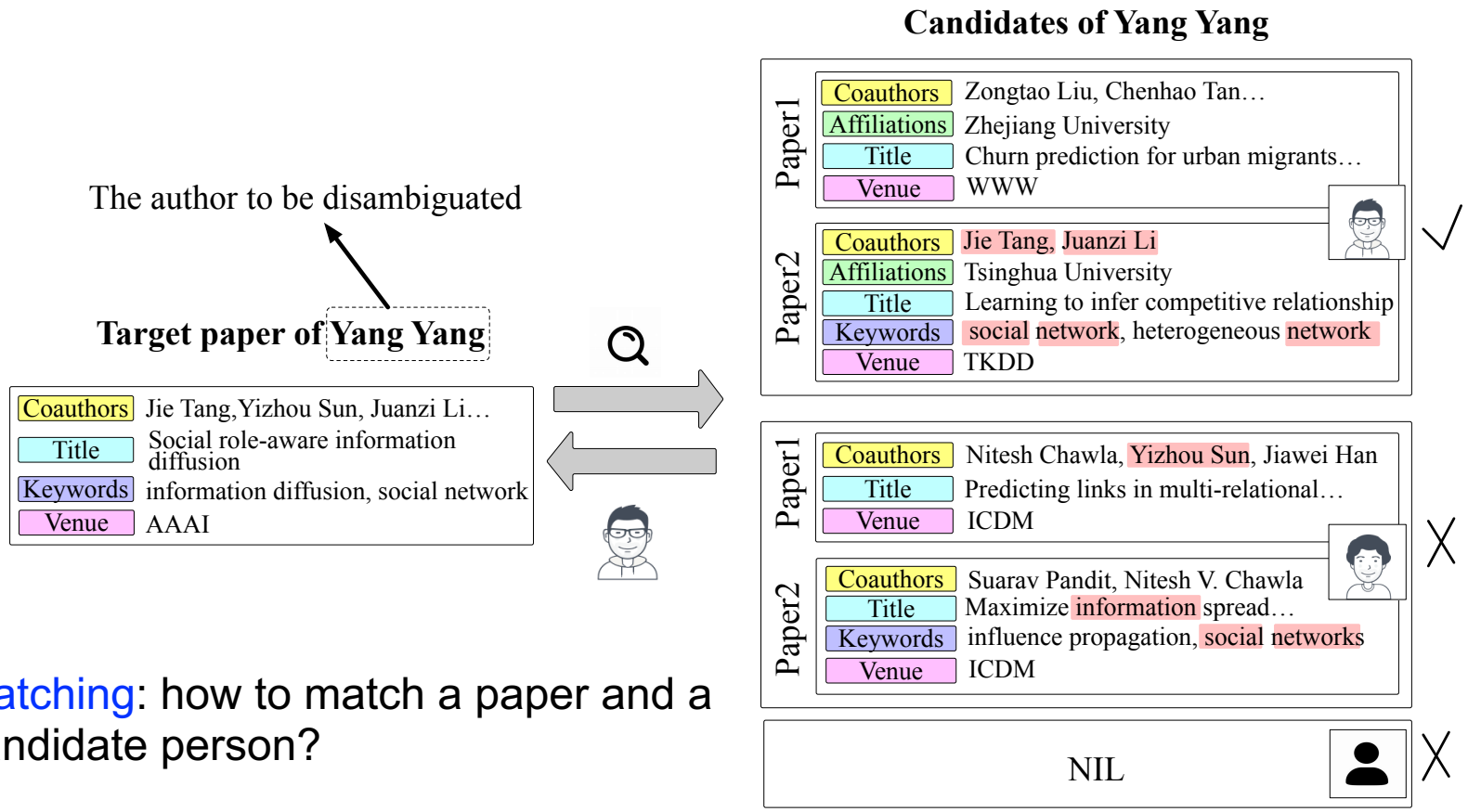
- Unassigned papers
  - New papers of AMiner > 50,000 per month





Bo Chen, Jing Zhang, Jie Tang. CONNA: Addressing Name Disambiguation on The Fly. Submitted to TKDE

# Challenges



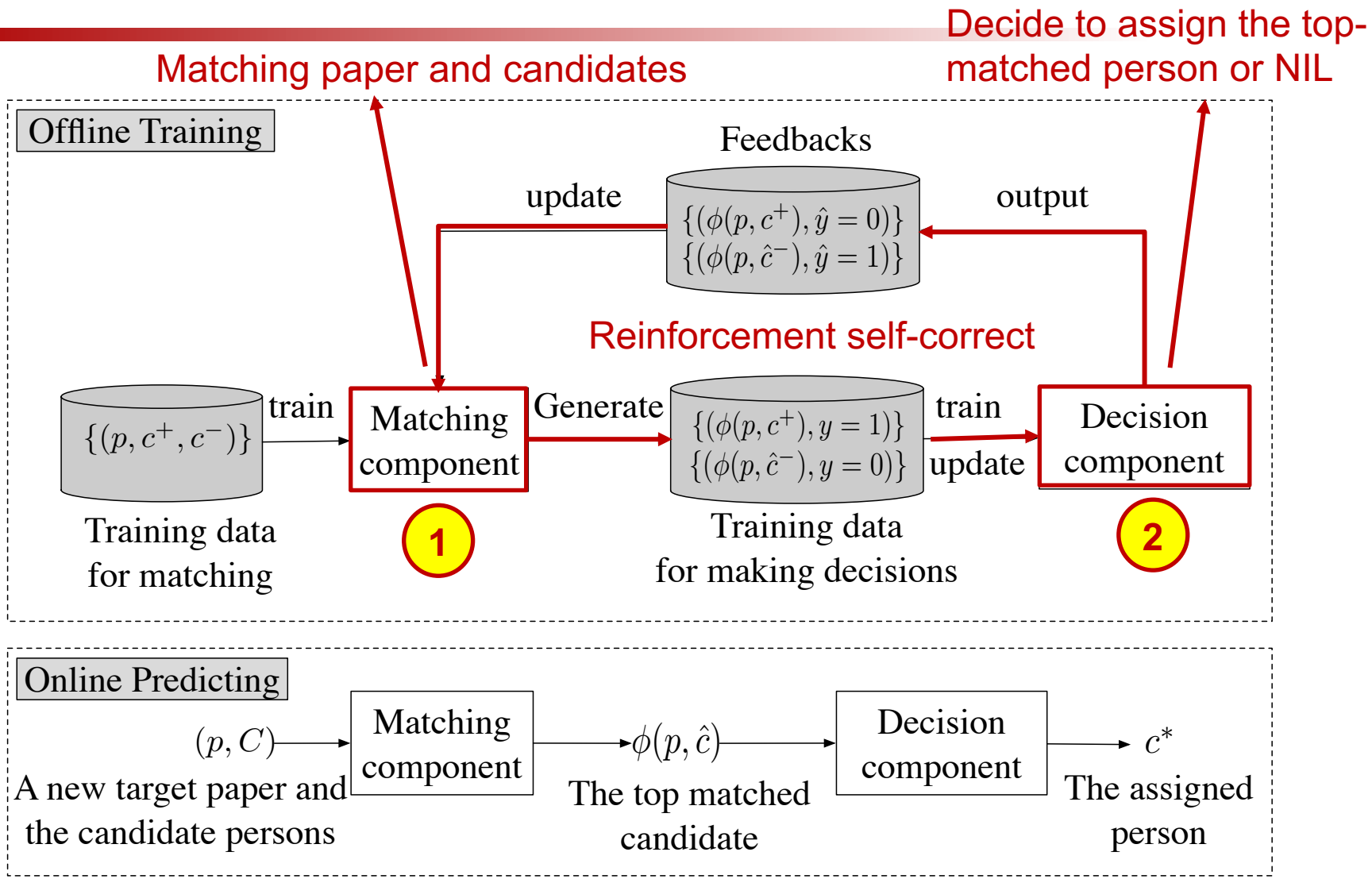

Matching: how to match a paper and a candidate person?

Decision: how to decide to assign the top matched candidate or NIL?

# Problem Definition

- Name disambiguation on the flye
  - Input: a target paper ***p*** and the set of candidate persons ***C***
  - Output: the right person $\boldsymbol{c}^*$ that ***p*** can be assigned, where $\boldsymbol{c}^*$ is a real person $\boldsymbol{c}^+$ or a non-existing person *NIL*.

# Framework

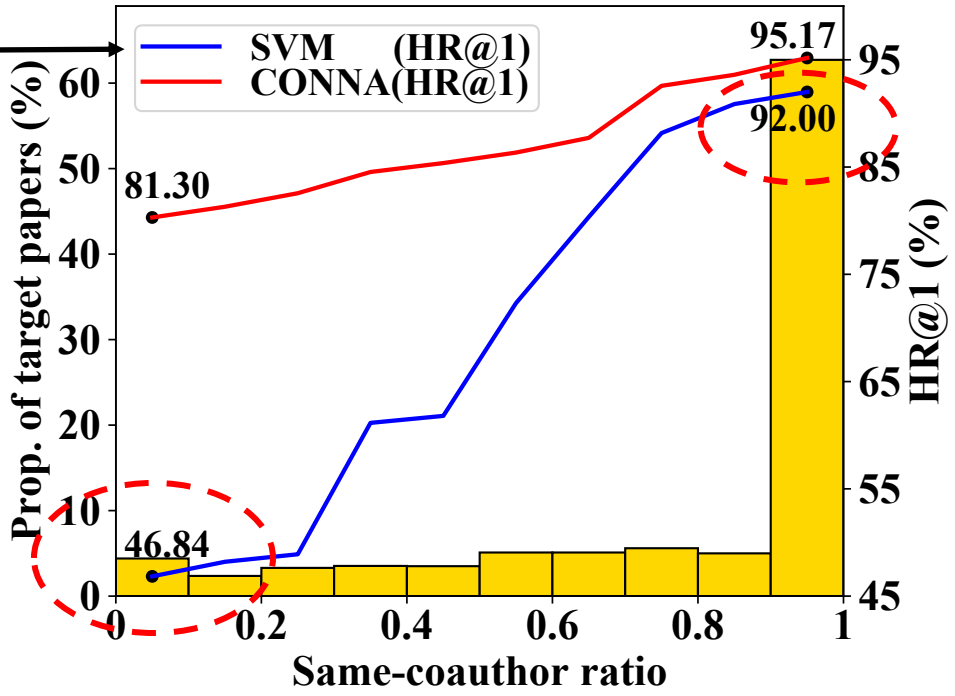

Matching paper and candidates
Decide to assign the top-matched person or NIL
Offline Training
Feedbacks
update
output
$\{(\phi(p,c^+),\hat{y}=0)\}$
$\{(\phi(p,\hat{c}^-),\hat{y}=1)\}$
Reinforcement self-correct
$\{(p,c^+,c^-)\}$
train
Matching component
Generate
$\{(\phi(p,c^+),y=1)\}$
$\{(\phi(p,\hat{c}^-),y=0)\}$
train
update
Decision component
Training data for matching
1
Training data for making decisions
2
Online Predicting
$(p,C)$
A new target paper and the candidate persons
Matching component
$\phi(p,\hat{c})$
The top matched candidate
Decision component
$c^*$
The assigned person

# Matching

- How can coauthor names take effect ?

**Name is important !** [WSDM, 2013]
56% same users with different accounts across the social networks can be correctly linked together

Feature-based

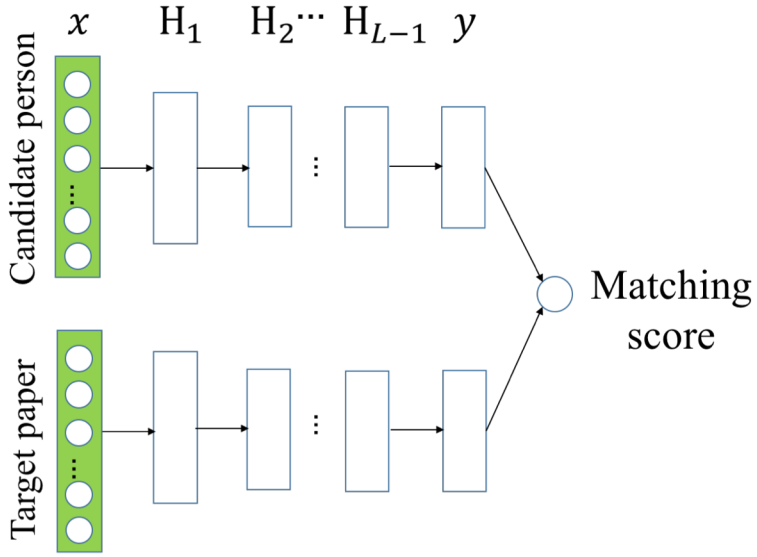



If there are many similar coauthors, it is easy to match

If there are only a few same coauthors, it is hard to match

# How to Improve the Matching Performance?

- Feature-based
  - Exact matching the tokens
- Representation-based
  - Semantic matching a paper and a person

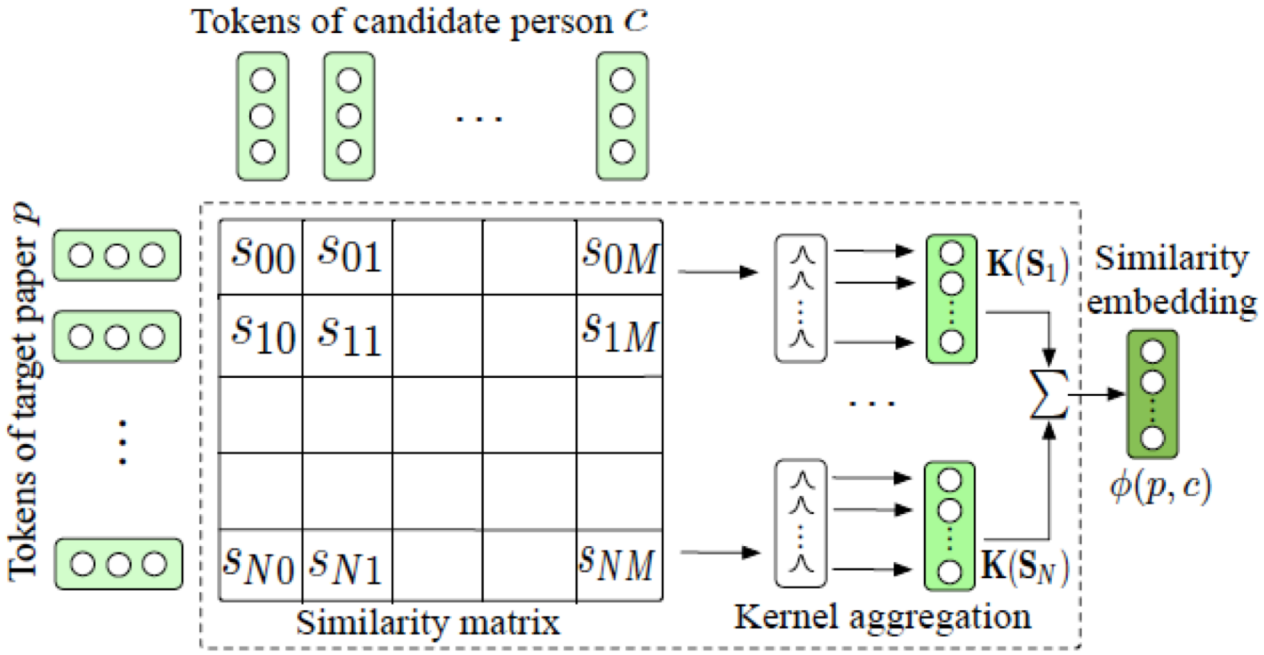



Dilute the effect of the exact matching.

E.g., exact matching is suitable for comparing coauthor names

# Basic Interaction Matching Model

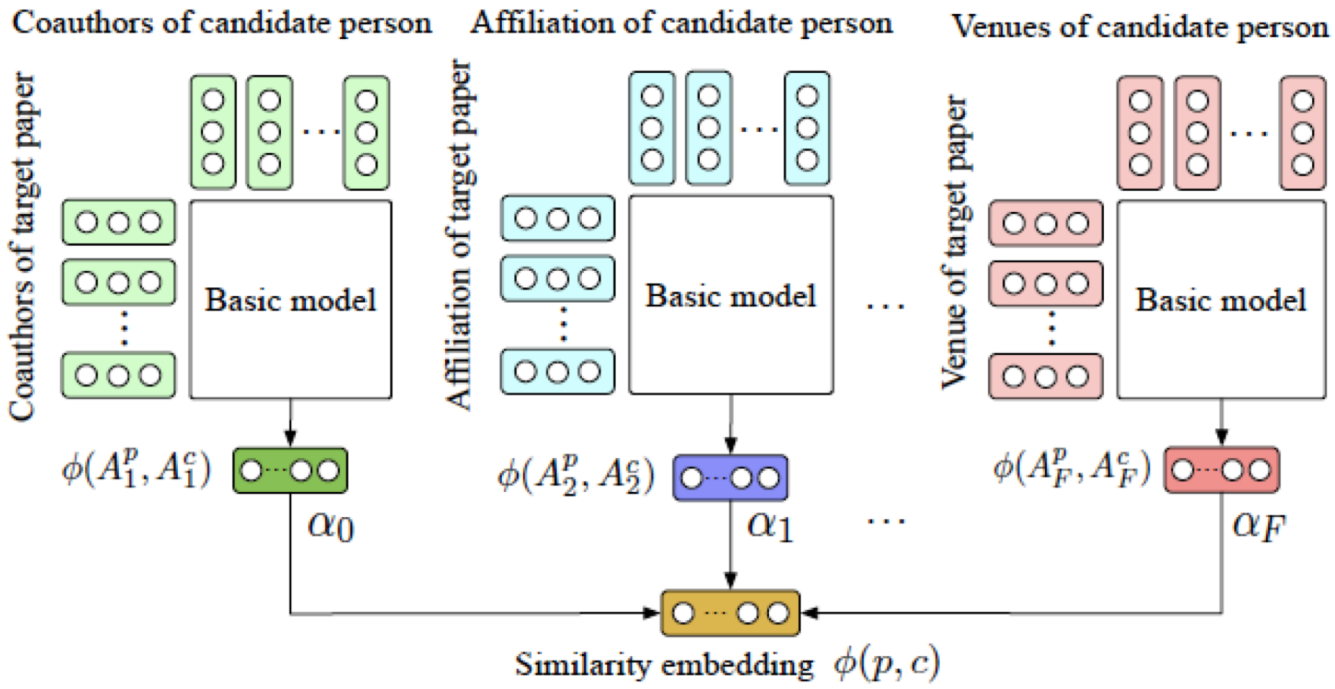
- Capture both the exact and soft matches




Xiong, et al. SIGIR'17

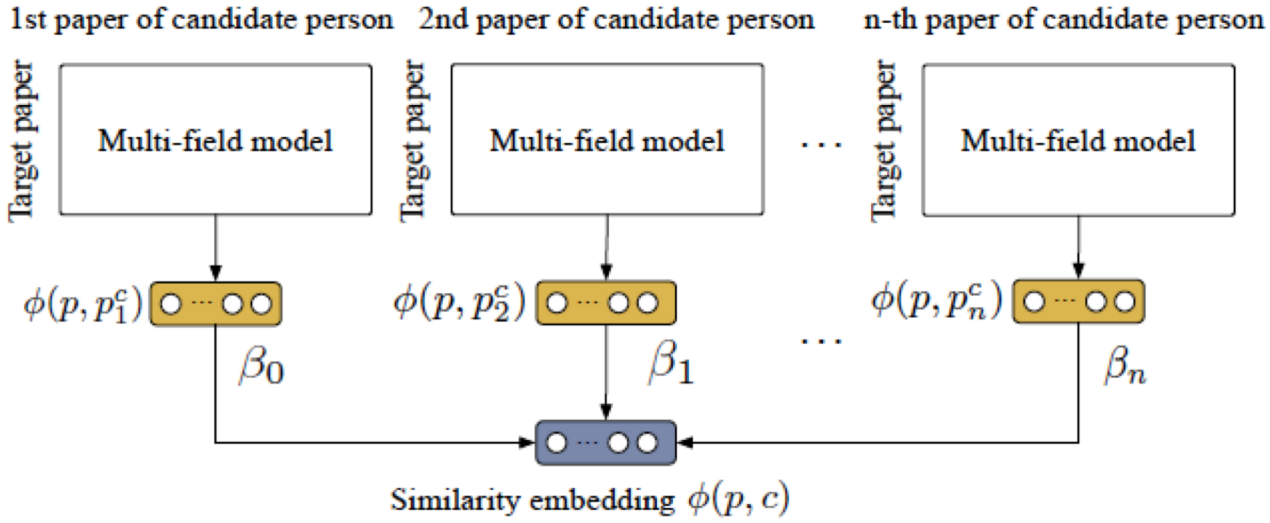
# Modeling Multi-field Attributes

- Assumption
  - Different fields of attributes takes different effects.



# Modeling Multi-Instances of Papers

- Assumption
  - Different papers of a candidate person takes different effects.



# Objective Function: Learning to Rank

- Triplet loss function:

$$\mathcal{L}(\Theta) = \sum_{(p,c^+,c^-)\in\mathcal{D}} \max\{0, g(\phi(p,c^+)) - g(\phi(p,c^-)) + m)\}.$$

  - $m$: margin between positive and negative pairs
  - $g$: transform feature vector $\phi$ to a score

# Decision

- Decide to assign the top-matched person (y=1) or NIL(y=0).

- Training data:
  - Positive instances: $\{(\phi(p, c^+), y = 1)\}$
  - Negative instances: $\{(\phi(p, c^-), y = 0)\}$

Top-matched person by the matching model

- Objective: train a classification model

$$h(\psi): \{\phi(p, c)\} \rightarrow \{y\}$$

# Reinforcement Self-Correction

- Generator:
  - Treat the matcher as a generator to generate the features $\phi(p, c)$
- Feedback:
  - The Decision classifier $h(\psi)$: $\{\phi(p, c)\} \rightarrow \{y\}$ give feedback to the matcher

$$R(y, \hat{y}) = \begin{cases} 1 & \hat{y} = y; \\ 0 & \text{otherwise.} \end{cases}$$

# Training Process

**Algorithm 1**: Reinforcement Joint Training

**Input**: A training set $\{(p, C)\}$.
**Output**: A matching component and a decision component parametrized by $\Theta$ and $\Phi$ respectively.

Construct $\mathcal{D}^r = \{(p, c^+, c^-)\}$;
Pre-train $\Theta$ of the matching component on $\mathcal{D}^r$;
Construct $\mathcal{D}^c = \{(\phi(p, c), y)\}$ by the matching component;
Pre-train $\Phi$ of the decision component on $\mathcal{D}^c$;
**repeat**
  **for** $(\phi(p, c), y) \in \mathcal{D}^c$ **do**
    Predict $\hat{y}$ by the decision component ;
    Calculate $R(y, \hat{y})$ by Eq. (9) ;
    Calculate $\nabla_\Theta J(\Theta)$ by Eq. (10);
    $\Theta \rightarrow \Theta + \mu \nabla_\Theta J(\Theta)$, where $\mu$ is the lea
  Regenerate $\mathcal{D}^c$ by the matching component;
  Update $\Phi$ in the decision component on $\mathcal{D}^c$;
**until** *Convergence*;

Pre-train two components

Reward: punish the wrong features, reward the right features

$$R(y, \hat{y}) = \begin{cases} 1 & \hat{y} = y; \\ 0 & \text{otherwise.} \end{cases}$$

Gradient

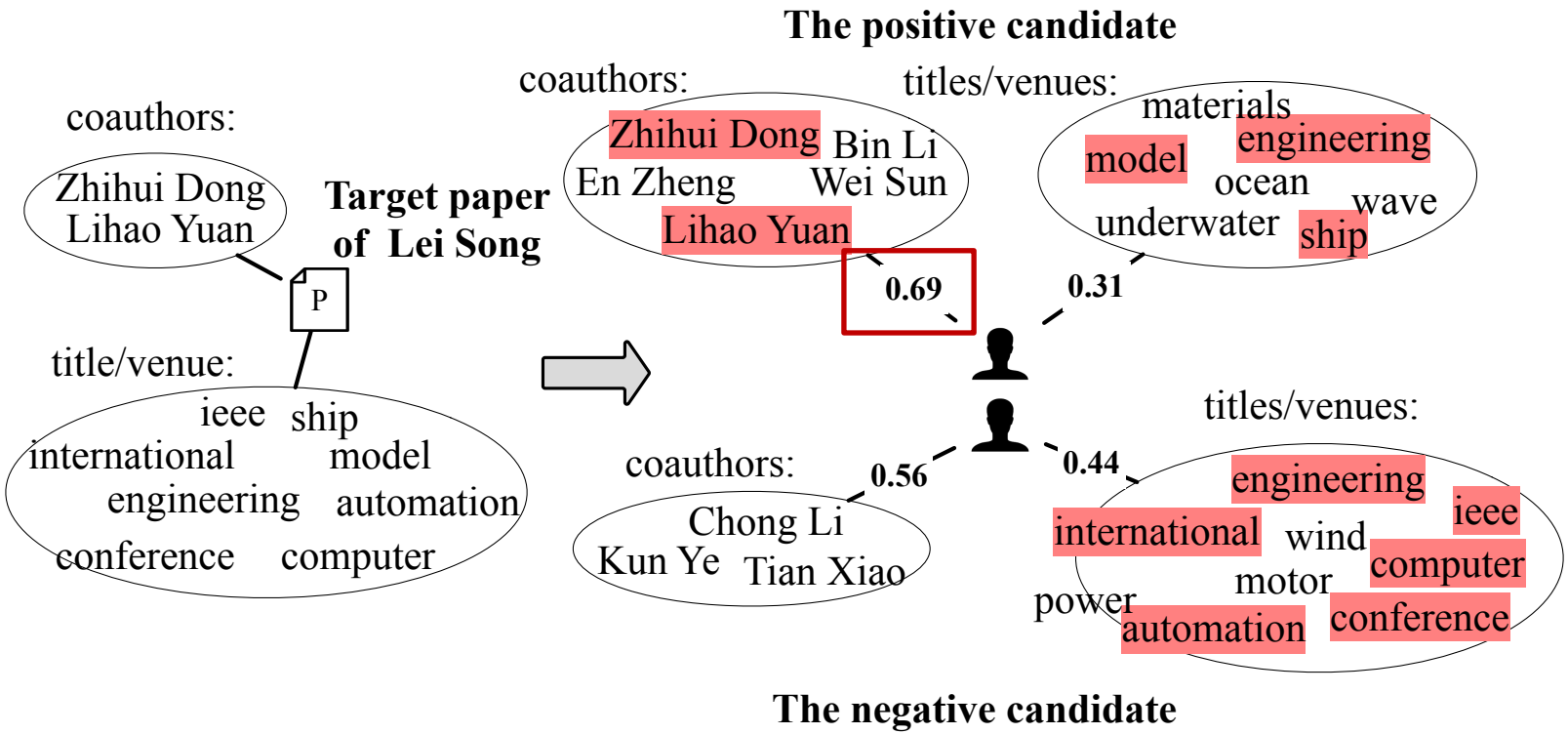
$$\nabla_\Theta J(\Theta) \sum_{(\phi(p,c),y) \in \mathcal{I}} R(y, \hat{y}) \nabla p_\Theta(\phi(p, c))$$

# Experimental Results

Performance of the matching results (%).

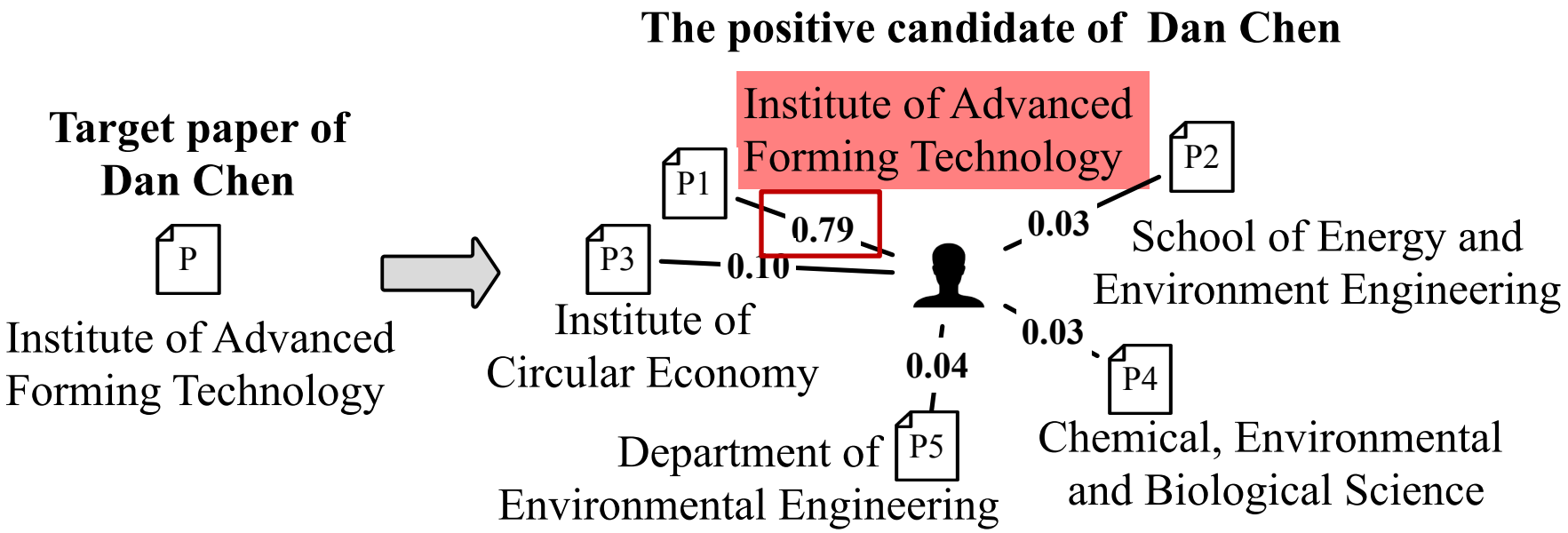
| Model | HR@1 | HR@3 | HR@5 | MRR |
|---|---|---|---|---|
| Camel | 68.02 | 89.27 | 93.96 | 79.40 |
| HetNetE | 75.85 | 93.74 | 97.05 | 85.26 |
| GML | 60.80 | 90.20 | 96.40 | 73.44 |
| SVM | 83.10 | 94.49 | 96.30 | 88.69 |
| CONNA$^r$(BP) | 86.20 | 96.40 | 98.41 | 92.20 |
| CONNA$^r$(MFP) | 88.00 | 98.75 | 99.50 | 93.25 |
| CONNA$^r$(MFMI) | 89.45 | 98.40 | 99.60 | 93.82 |
| **CONNA** | 90.70 | 98.50 | 99.40 | 94.59 |
| **CONNA+Fine-tune** | **91.71** | **98.90** | **99.60** | **95.24** |

# Effect of Multi-field Modeling




The field of coauthors is more important than other fields

# Effect of Multi-Instance Modeling



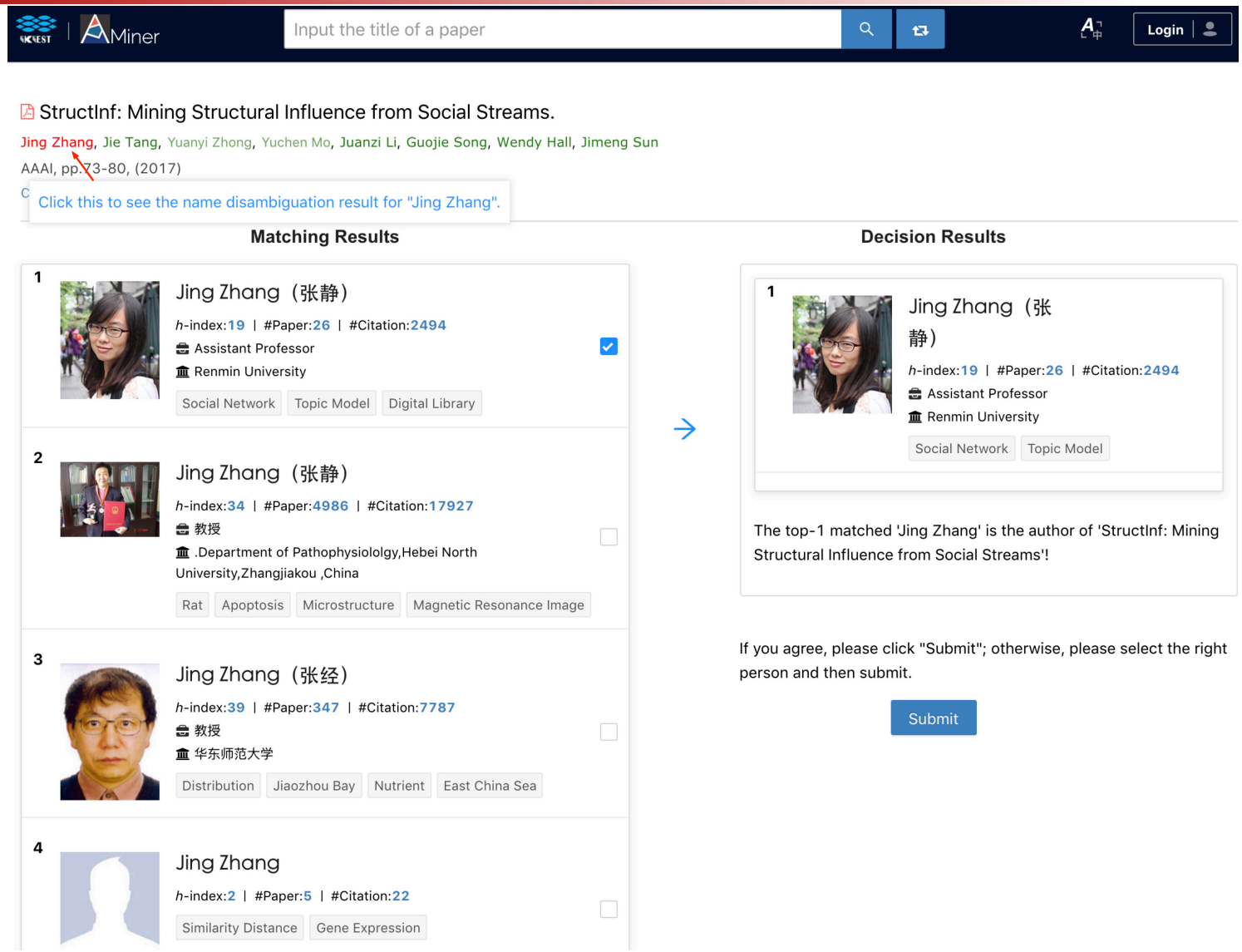

Paper 1 is more important than other papers

# Experimental Results

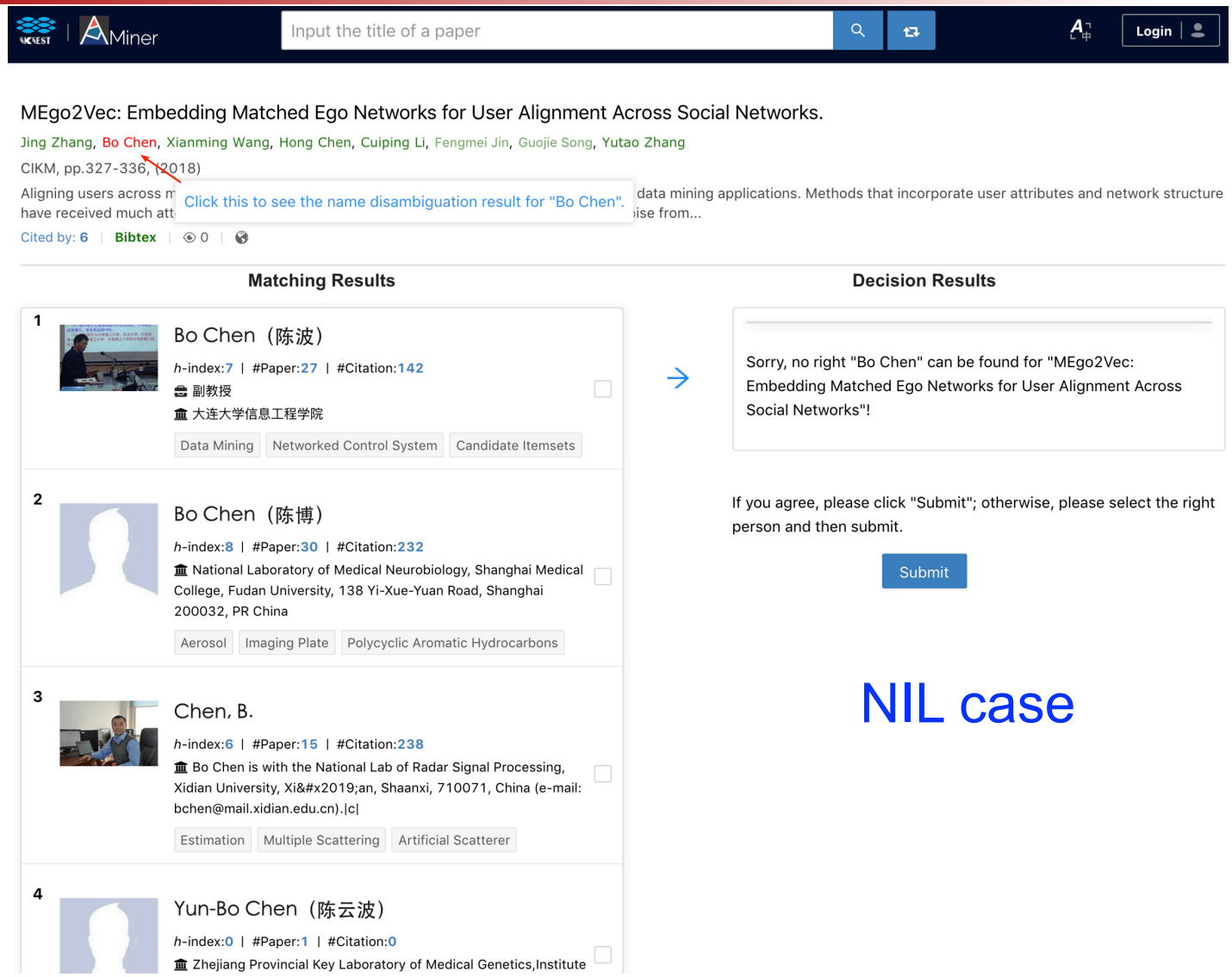
Performance of the decision results (%).

| Model | Samples with $c^* = c^+$ | | | Samples with $c^* = \text{NIL}$ | | |
|---|---|---|---|---|---|---|
| | Pre. | Rec. | F1 | Pre. | Rec. | F1 |
| SVM | 79.47 | 66.60 | 74.47 | 71.26 | 82.80 | 76.60 |
| Threshold | 79.33 | 57.60 | 66.38 | 66.47 | **84.07** | 74.24 |
| Heuristic Loss | 71.79 | 78.40 | 74.95 | 76.21 | 69.20 | 72.54 |
| CrossEntropy | 79.42 | 82.33 | 80.85 | 81.66 | 78.67 | 80.14 |
| **CONNA** | 79.53 | 89.87 | 84.38 | 88.35 | 76.87 | 82.21 |
| **CONNA+Fine-tune** | **82.47** | **90.33** | **86.22** | **89.31** | 80.80 | **84.84** |

# Demo of Assigning Papers



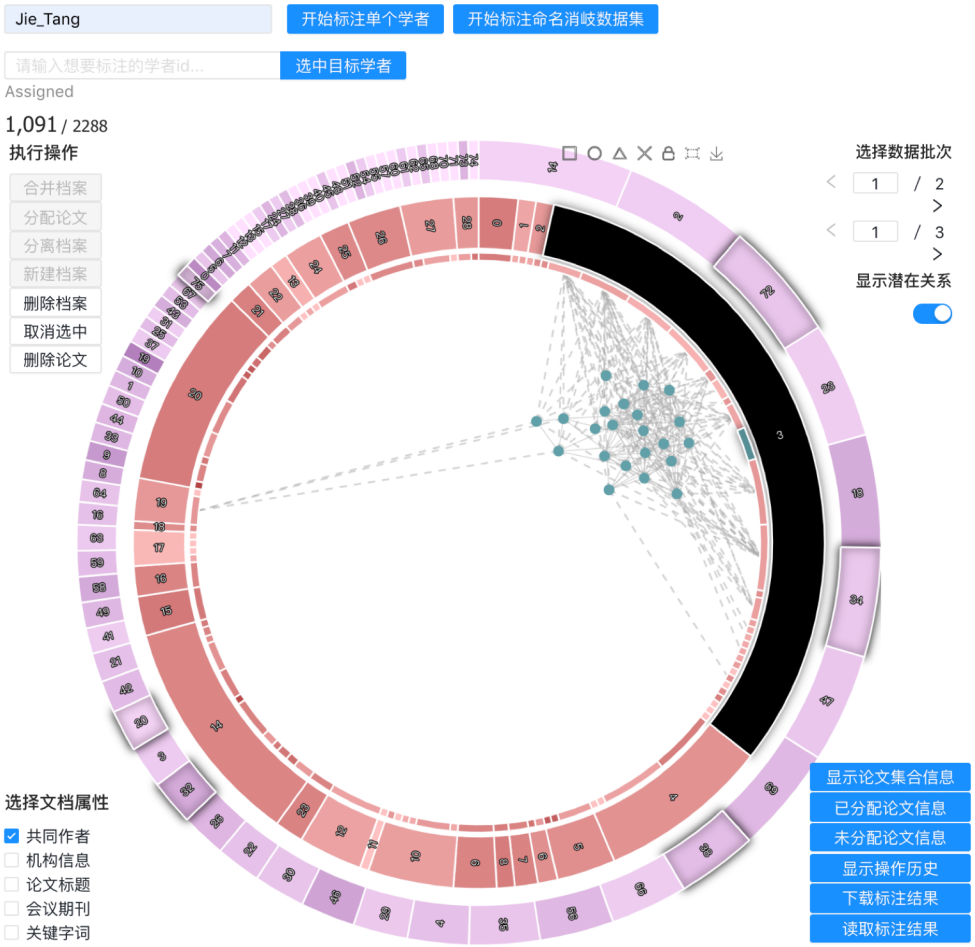
# Demo of Assign Papers



# Conclusion

- We are the first to study the problem of name disambiguation on the fly.
- We propose a multi-field multi-instance interaction-based model to match a paper and a person.
- We jointly train the matching and the decision component to boost the performance.

# Wrongly Assigned Papers




To deal with the over-partitioned and over-merged persons


How to recommend the most possible operations?



# Revisiting Name Disambiguation

- Focus on unassigned and wrongly assigned papers

Dataset: https://www.aminer.cn/diambiguate-data

Demo: http://na-demo.aminer.cn/

Challenge: https://biendata.com/competition/aminer2019/

**Thank You!**

Collaborate with Bo Chen (*RUC*), Jie Tang (*THU*)