Isnad Disambiguation

By Joe Hilleary and Kyle Sayers

Reframing the Problem

- Historians are interested in scholarly transmission
  - Captured by *isnads*
- Name disambiguation problem
  - Manually time consuming
- Can we make suggestions based on limited manual labeling to speed up the process?

حدثنا أبو داود قال :حدثنا هشام، عن قتادة، عن الحسن عن سمرة، أن النبي صلى الله عليه وسلم

Abū Dāwūd transmitted to us, saying, ‘Hishām transmitted to us, from Qatādah, from al-Ḥasan, from Samurah that the Prophet, may the peace and blessing of God be on him[1]

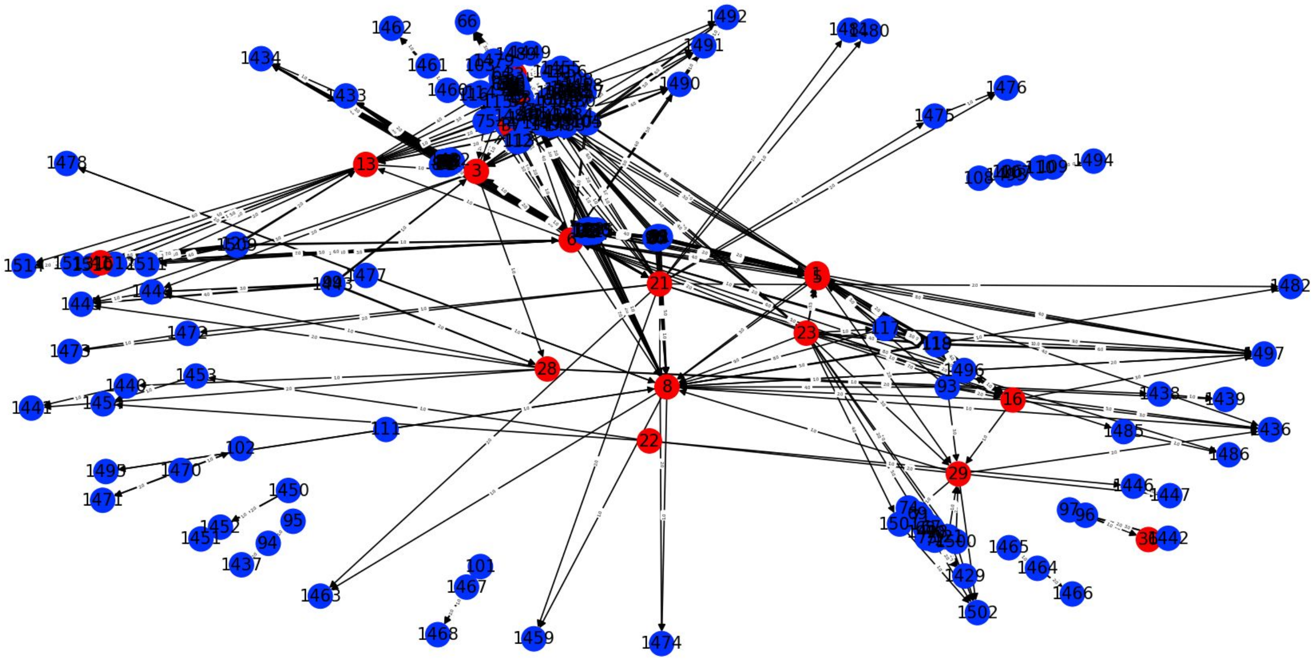
Starting Data

- Partially disambiguated chains
  - From Ta’rikh Madinat Dimashq by Ibn ‘Asakir
  - All connecting through Muhammed Ibn Sa’d
- 2,380 chains
- 14,454 mentions
  - 13,072 labeled by domain expert
    - 44 unique individuals

1. Building the Graph

Graph at t_0

- Read in the data
  - Select some labels to cover up
  - Connect nodes based on co-occurrences
  - Weight directed edges by average position

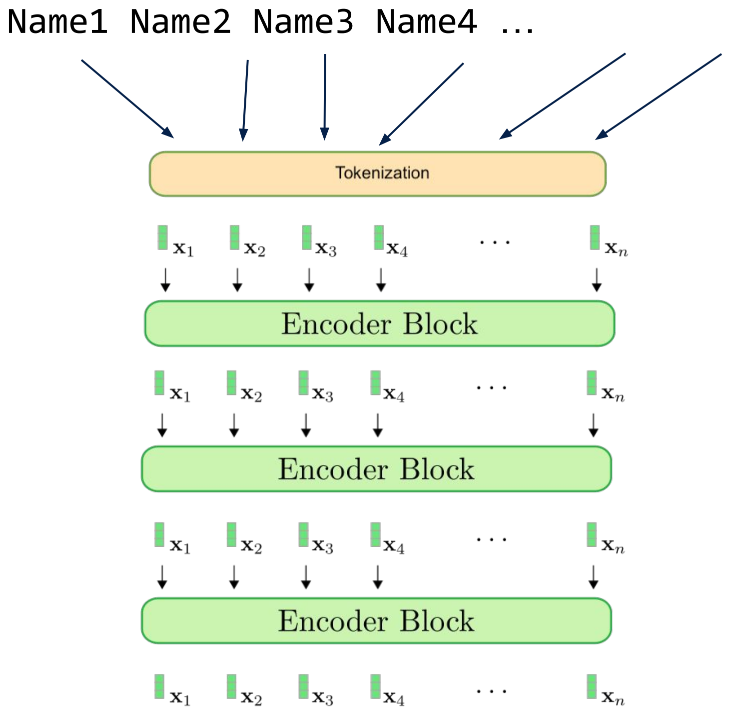

1492
1491
1490
1480
1462
66
1434
1461
1433
1476
1475
1494
1478
13
3
112
1509
1514
1477
1444
1443
1482
21
23
1497
1496
16
93
1438
1439
1436
1485
1486
1473
1472
28
8
1453
1441
111
22
29
102
1470
1471
1450
1452
1451
95
94
1437
1463
101
1467
1468
1459
1474
1446
1447
1442
1465
1464
1466
1429
1502

2. Calculating Features

Social Hashing

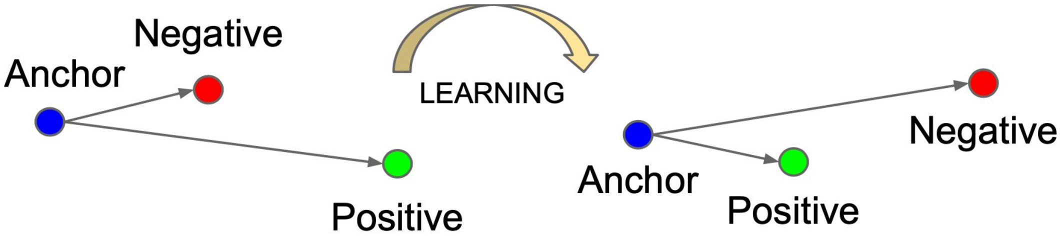
- 2 social feature vectors
  - # co-occurrences
  - average relative position in co-occurrences
- Length equal to the number of nodes in the graph

Context Embedded Tokens



Contrastive Name Embeddings

- Supervised learning of name representations
- Initially used the same test set for contrastive learning and graph prediction, but found it was unnecessary

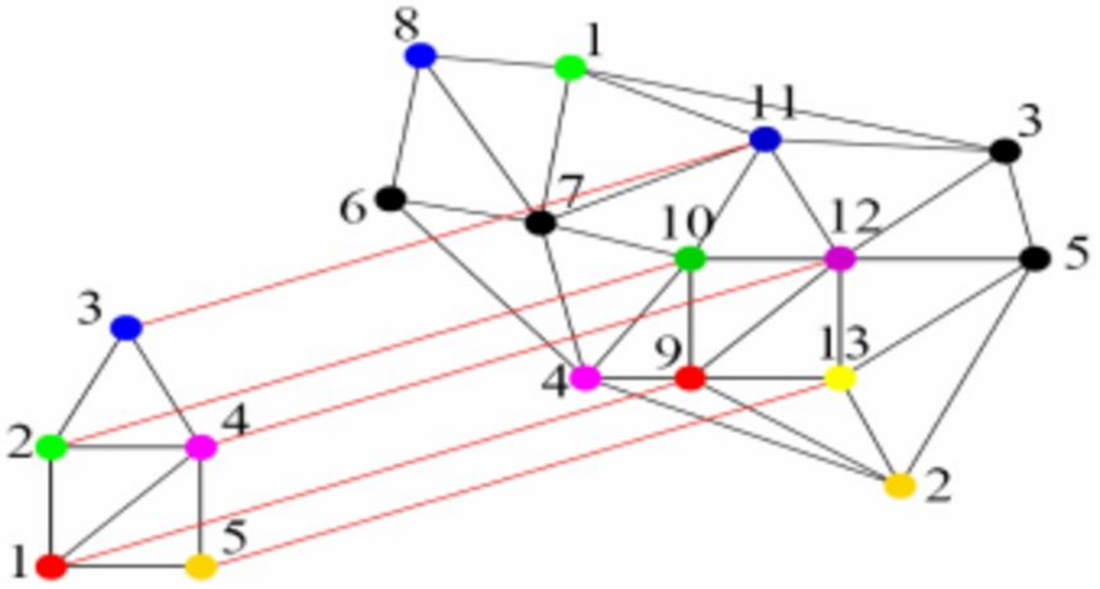


3. Match and Merge

Calculate Similarity Matrix

- Cosine similarity between all nodes
- Dynamic computation
- Memoization of results
- Multithreaded computation (limited)

Subgraph Matching



Merge Until Stable

- Each unlabeled node matched to label
  - Uniquely assigned
  - Assigned to an existing label
- Recompute features each iteration
  - Using features from composite nodes
- Stop when all ambiguous nodes have been assigned a label

4. Concessions for Performance

Computational Limits

- Size of the similarity matrix
- Networkx limits multithreading (graph tool)
- Recomputation after every merge

- More efficient hash recomputation
- Locality sensitive hashing

Processing

- Batches of top 50 nodes at each stage of recalculation
- Can’t match unlabeled node to unlabeled node
  - If nodes begin to resemble each other later on, matches are already fixed
- Only check labeled neighbors

5. Results

Evaluating

- CoNLL
  - Composite measure of cluster similarity
    - MUC, B^3, $CEAF_e$
- Did we re-cluster nodes known to share a label?

Comparison with Muther & Smith

	kNN100_leiden	Surface Form	Our Method (33%)
B^3	**.756**	**.868**	**.603**
$CEAF_e$	**.444**	**.523**	**.664**
CoNLL	**.727**	**.790**	**.753**

Our Method

	Muther & Smith Embeddings (33%)	Contrastive Embeddings (33%)	No Social Features (33%)
B^3	**.603**	**.742**	**.979**
$CEAF_e$	**.664**	**.530**	**.826**
CoNLL	**.753**	**.755**	**.934**

Conclusion

- NLP embeddings seem to be a better way to tackle this problem than social features
- Contrastive embeddings appear more effective

5. What’s Next?

This Project

- Parameter tuning
- Ambiguous to ambiguous matching
  - No known starting labels
- Shifting weights
  - Move from NLP to social as more is known

Future Work

- Locality sensitive hashing
  - Similarity -> collisions
- Deep features
  - *Neural Subgraph Matching (Rex Ying, Andrew Wang)*
- Jaccard Index for subgraph matching
- Other distance metrics for social features

Citations

Goebel, Peter & Vincze, Markus. (2007). Implicit Modeling of Object Topology with Guidance from Temporal View Attention.

R. Muther and D. Smith, ‘The Fellowship of the Authors: Disambiguating Names from Social Network Context’. arXiv, 2022.

R. Muther, D. Smith, and S. Savant, ‘From Networks to Named Entities and Back Again Exploring Classical Arabic Isnad Networks’. *Journal of Historical Network Research* 5, 2023.

Rex *et al.*, ‘Neural Subgraph Matching’. arXiv, 2020.