Medical Entity Disambiguation Using Graph Neural Networks

Alina Vretinaris	IBM Germany
Chuan Lei	IBM Research – Almaden
Vasilis Efthymiou	FORTH – ICS
Xiao Qin	IBM Research – Almaden
Fatma Özcan	Google

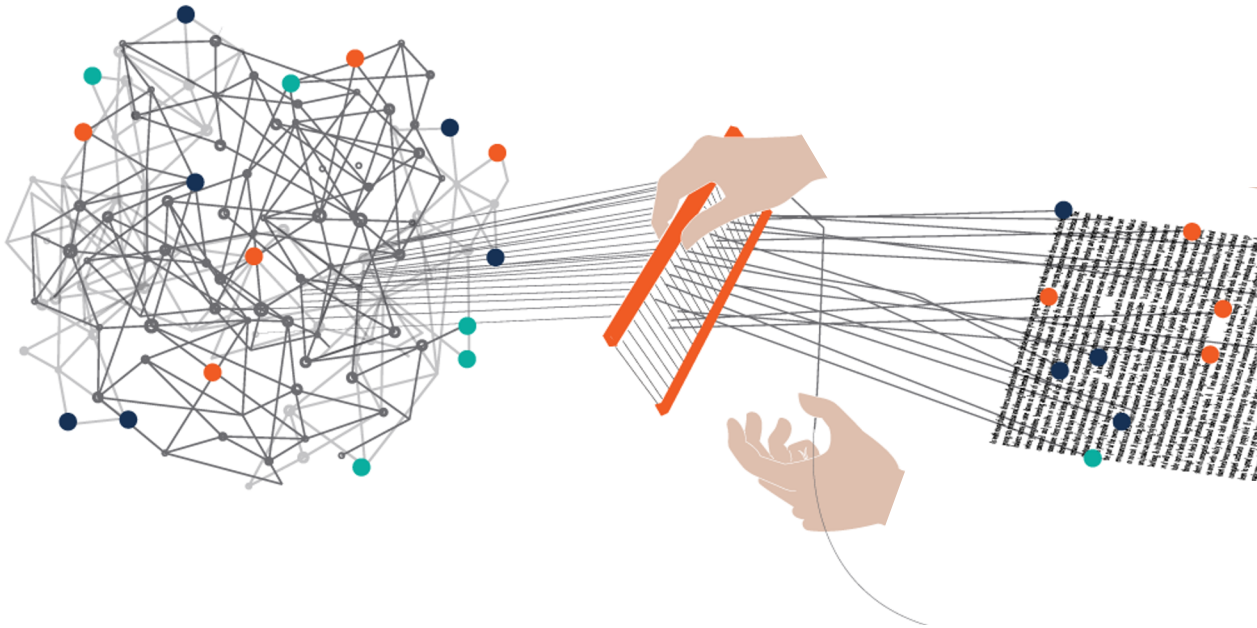


Photo credit: Ontotext

Medical Entity Disambiguation

- Medical knowledge base (ontology) curation/maintenance
- Editorial staff often refer to a medical entity (a.k.a concept/class) in a knowledge base with acronyms, typos and colloquial terms
- Example
  - An entity in KB: *acute renal failure*
  - Text snippet from an editorial staff: *Aspirin can cause nausea, indicating a potential ARF, nephrotoxicity, or proteinuria.*
  - *ARF* is the ambiguous term
  - Should *ARF* be resolved to *acute renal failure* or *acute rheumatic fever*?

Discrepancies arises during the curation/maintenance process, needing entity disambiguation capability

Main Challenges

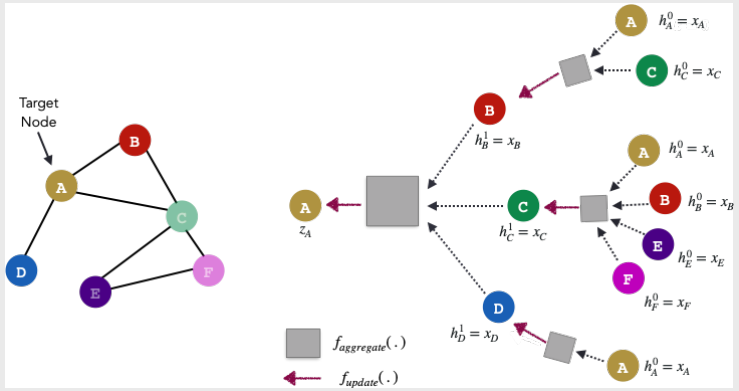
- Semantic gap between a text snippet and a medical KB incurred by discrepancies

- Collectively learn contextual and structural information of entities in a text snippet
  - Ambiguous terms in a text snippet should not be considered in isolation
  - Infuse domain knowledge with the recognized entities in a text snippet

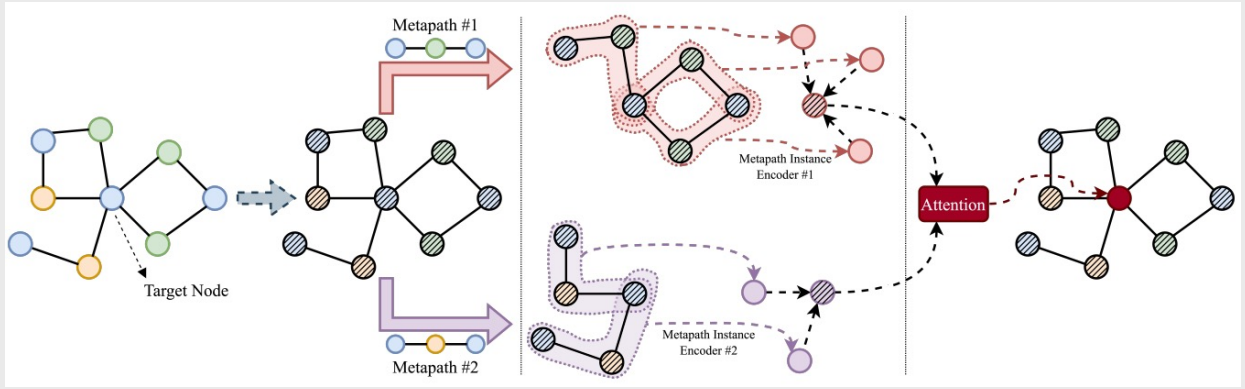
- Capture discriminative contextual information of entities in a medical KB
  - Not all neighbors of a concept in a medical KB are equally important to the ambiguous term
  - Learn from difficult samples to improve the capability of disambiguation model

Three Representative Graph Neural Networks

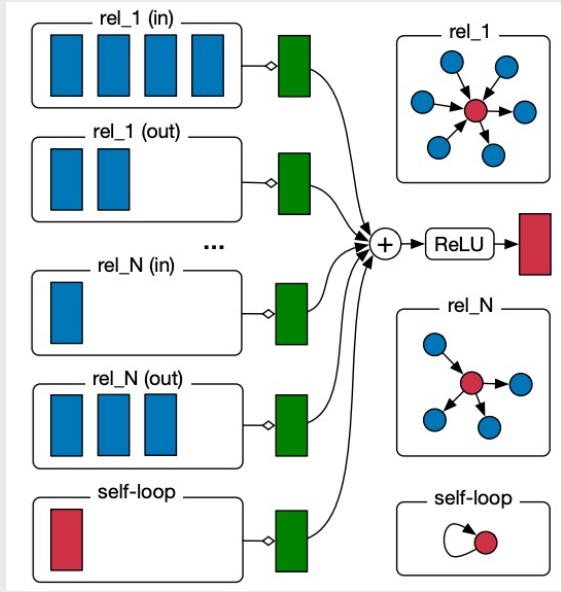
GraphSAGE [1] – seminal message-passing GNN



MAGNN [3] - metapath aggregated GNN for heterogeneous graphs

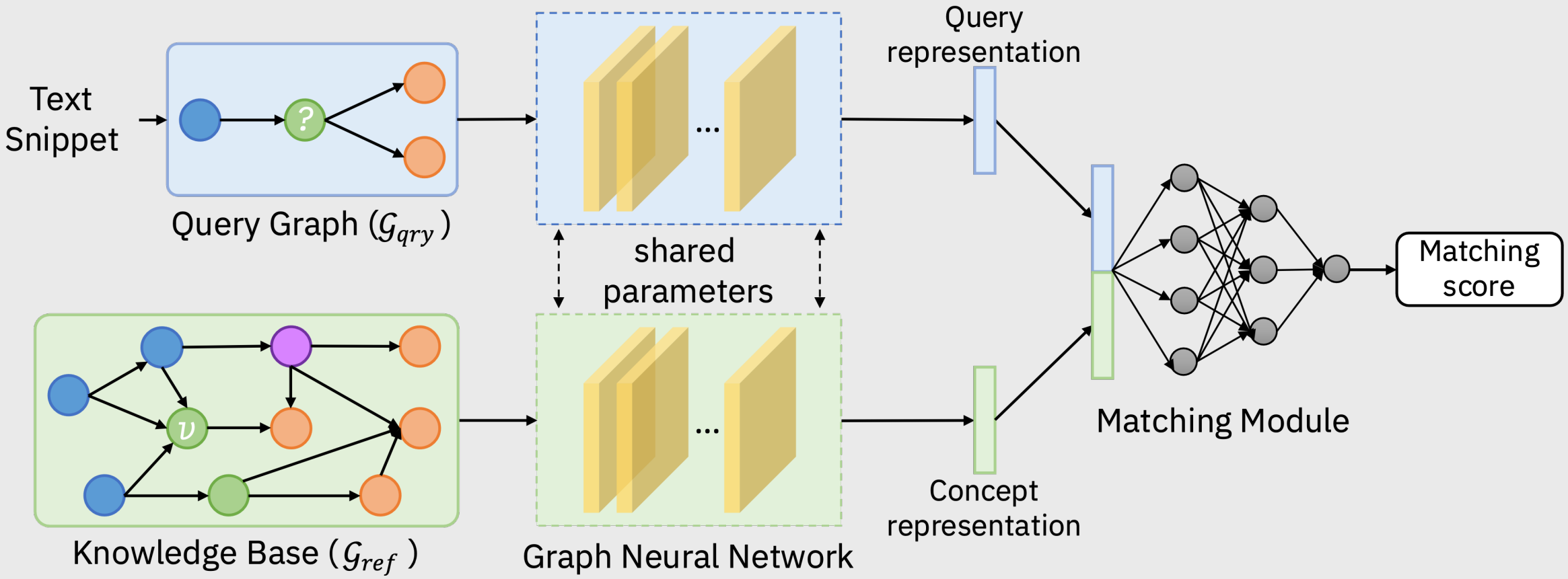


R-GCN [2] – relation-aware GNN, distinguishing different neighbors with specific relations



1. W. L. Hamilton, R. Ying, and J. Leskovec. Inductive representation learning on large graphs. In NIPS, pages 1024–1034, 2017.
2. M. S. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling. Modeling relational data with graph convolutional networks. In ESWC, pages593–607, 2018.
3. X. Fu, J. Zhang, Z. Meng, and I. King. Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding. In WWW, page 2331–2341, 2020.

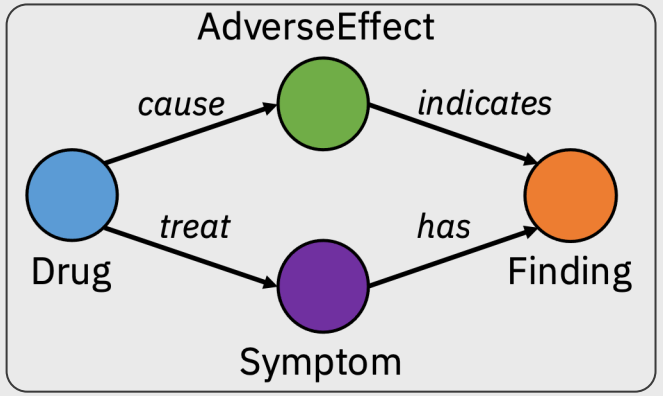
ED-GNN Architecture



Semantic Augmentation for Query Graph

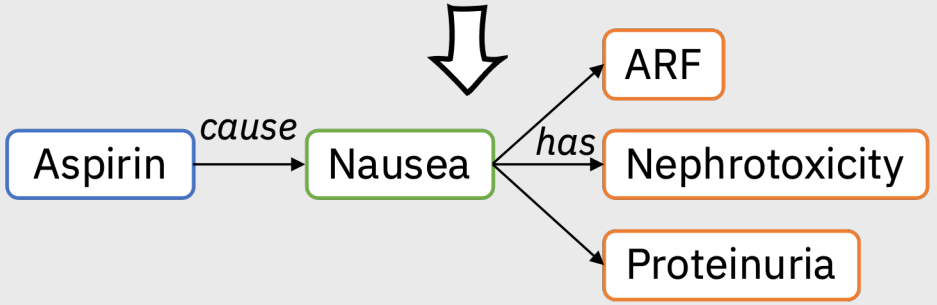
- Augment entity mentions with node types from the KB (G_{ref})
  - Drug – Aspirin
  - AdverseEffect – Nausea
  - Finding – ARF, Nephrotoxicity, proteinuria
- Augment relationships in the query graph (G_{qry})
  - Aspirin-*cause*->Nausea
  - Nausea-*has*->ARF
  - Nausea-*has*->Nephrotoxicity
  - Nausea-*has*->Proteinuria

KB Schema



Text Snippet

Aspirin can cause nausea indicating a potential ARF, nephrotoxicity, or proteinuria.



Query Graph (G_{qry})

Semantic-Driven Negative Sampling

- Negative sampling
  - Sampling difficult negative triplets avoids the problem of vanishing gradient and thus obtain better performance.
- Semantic similarity sim_{se} (negative samples semantically similar to the positive one)
  - Positive pair – (MH, Malignant hyperpyrexia)
  - Negative pair – (MH, Malignant hyperthermia)
- Structural similarity sim_{st} (negative samples share many common neighbors with the positive one)
  - Graph similarity metrics – graph edit distance (GED), maximum common subgraph, graph kernels
  - Limit to 1-hop neighbor to reduce the computational cost
- Scoring function – $sim_{se} \cdot sim_{st}$

Experimental Setup

- Datasets
  - MDX – a medical KB about drugs, adverse effects, indications, findings, etc.
  - MIMIC-III – public data set with 40,000 anonymized patient health-related records
  - Bio CDR – 1,500 PubMed abstracts annotated with mentions of chemicals, diseases
  - NCBI – 700 PubMed abstracts annotated with disease mentions and their corresponding concepts in MeSH
  - ShARe – 433 anonymized clinical notes from MIMIC II clinical dataset, annotated with disorder mentions
- Baselines
  - ED-GNN variants – ED-GNN (GraphSAGE), ED-GNN (R-GCN) and ED-GNN (MAGNN)
  - DeepMatcher – a supervised deep learning solution designed for entity resolution in a tabular setting
  - NormCo - a deep coherence model for disease entity normalization
  - NCEL - creates a graph for candidates of mentions and then apply GCN to improve the disambiguation by directly aggregating information from linked nodes

Experimental Results

- Overall results
  ED-GNN variants consistently outperform other solutions in terms of precision, recall, and F1 on all datasets

Methods	DeepMatcher			NormCo			NCEL			ED-GNN (GraphSage)			ED-GNN (R-GCN)			ED-GNN (MAGNN)		
Datasets	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
MDX	0.656	0.700	0.677	0.687	0.634	0.659	0.673	0.659	0.666	0.614	0.900	0.730	0.722	0.867	0.788	**0.725**	**0.967**	**0.829**
MIMIC-III	0.708	0.567	0.630	0.747	0.692	0.718	0.716	0.624	0.667	0.786	**0.733**	**0.759**	0.810	0.567	0.667	**0.826**	0.633	0.717
NCBI	0.783	0.815	0.799	0.863	0.818	0.840	0.816	0.793	0.804	**0.924**	**0.860**	**0.891**	0.912	0.823	0.865	0.915	0.861	0.887
ShARe	0.694	0.639	0.665	0.726	0.623	0.671	0.753	0.631	0.687	0.794	0.829	0.811	0.806	0.833	0.819	**0.824**	**0.875**	**0.851**
Bio CDR	0.837	0.816	0.826	0.866	0.805	0.834	0.857	0.829	0.843	0.853	0.845	0.849	**0.896**	**0.866**	**0.881**	0.864	0.853	0.858

- Ablation studies
  Both query graph augmentation and semantic-driven negative sampling improve the basic ED-GNN substantially over different datasets

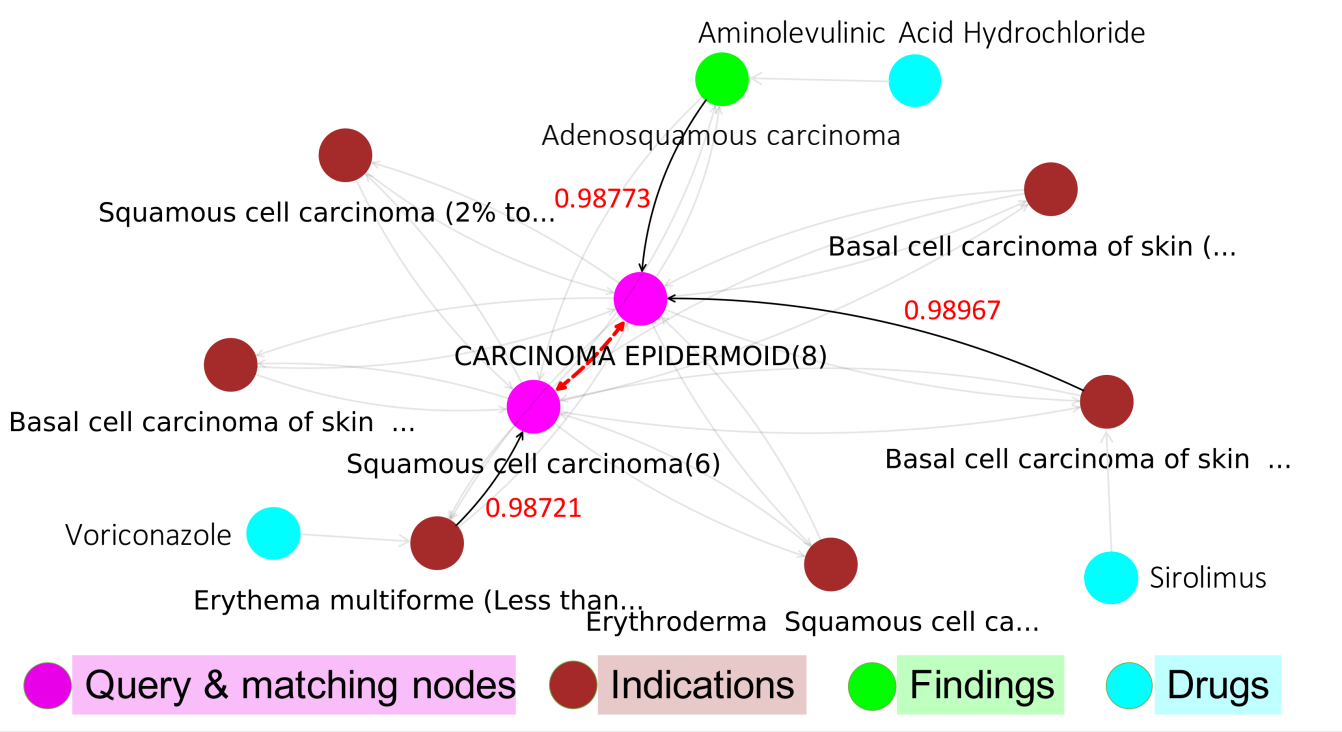
Methods	Datasets	Basic			Query graph augmentation			Negative sampling		
		Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
ED-GNN (GraphSage)	MIMIC-III	0.747	0.702	0.724	0.747	0.702	0.724	**0.786**	**0.733**	**0.759**
	NCBI	0.869	0.821	0.844	0.869	0.821	0.844	**0.924**	**0.856**	**0.889**
ED-GNN (R-GCN)	Bio CDR	0.825	0.798	0.811	**0.863**	**0.826**	**0.844**	0.846	0.805	0.825
ED-GNN (MAGNN)	MDX	0.671	0.827	0.741	0.694	0.863	0.769	**0.713**	**0.925**	**0.805**
	ShARe	0.754	0.824	0.787	0.796	**0.868**	**0.830**	**0.813**	0.842	0.827

- Number of layers in GNN – 2/3 layers provide the best results

# layers	MDX	MIMIC-III	NCBI	ShARe	Bio CDR
1	0.691	0.641	0.815	0.731	0.785
2	0.751	0.704	**0.891**	0.825	0.843
3	**0.829**	**0.759**	0.867	**0.851**	**0.881**
4	0.743	0.727	0.831	0.806	0.829

Case Study

- Highlight 3 most important edges that contribute to matching *squamous cell carcinoma* with *carcinoma epidermoid*, carrying information from different types of neighboring nodes
- ED-GNN learns and leverages the most semantically and structurally meaningful information among different types of entities and relations for entity disambiguation



Conclusions

- Model medical ED as a graph matching problem to leverage GNNs with a simple architecture
  Introduce ED-GNN, a GNN-based medical ED solution, employing GraphSAGE, R-GCN and MAGNN

- Two optimization techniques to improve ED-GNN's disambiguation capability
  - Construct the query graph and augment it with domain knowledge from the medical KB
    Help ED-GNN focus on the right structural information from the query graph for making the matching decisions
  - Design an effective negative sampling strategy, which provides ED-GNN with harder examples, resulting in more discriminative power for entity disambiguation

- Evaluate the effectiveness of ED-GNN on multiple real-world datasets
  Experimental results show ED-GNN consistently outperforms the state-of-the-art ED solutions in all datasets by up to 16.4% in F1 score

Thank you!

Questions?