Medical Entity Disambiguation Using Graph Neural Networks

Vretinaris et al.

(SIGMOD'21)

Presentation by Yannis Marketakis

In the context of HY563 (Computer Science Department, UOC)

Outline

In one slide

- Existing Medical Entity Disambiguation (ED) approaches are not adequate due to word discrepancies

- ED-GNN is a medical Entity Disambiguation system based on GNNs

- Entity Disambiguation is modeled as a **graph matching problem**

- Optimized features
  - Query graph augmentation based on medical KB
  - Negative sampling to improve disambiguation capability

- Implemented on top of GNNS
  - GraphSAGE, R-GCN, MAGNN

- Outperforms state-of-the-art solutions
  - Avg. Improvement of 7.3 % on F1-score over real-world datasets

GNN
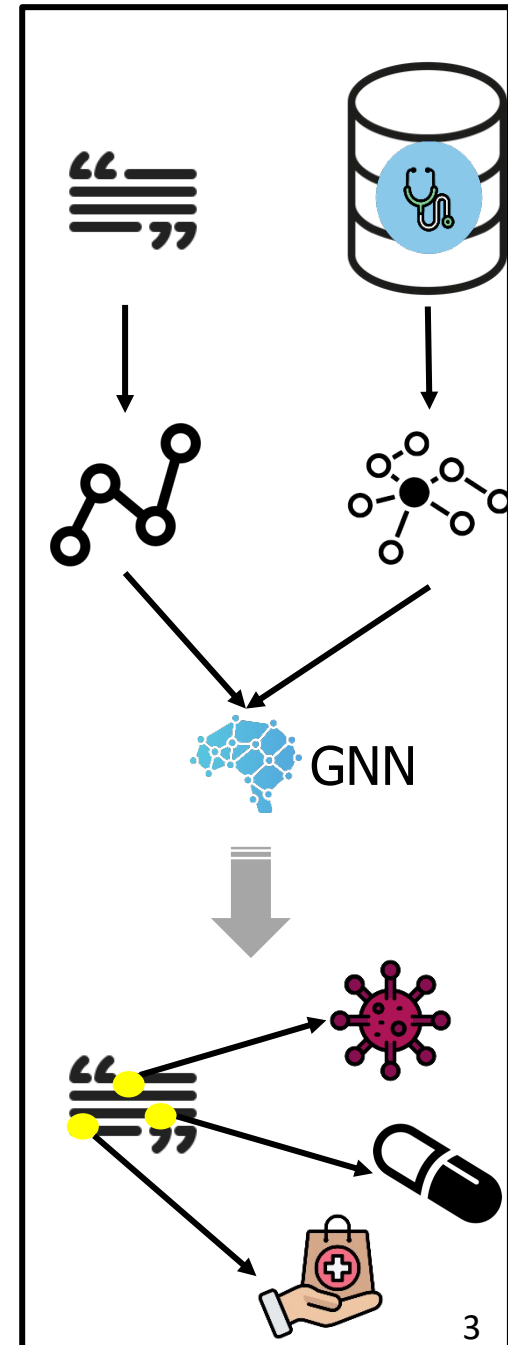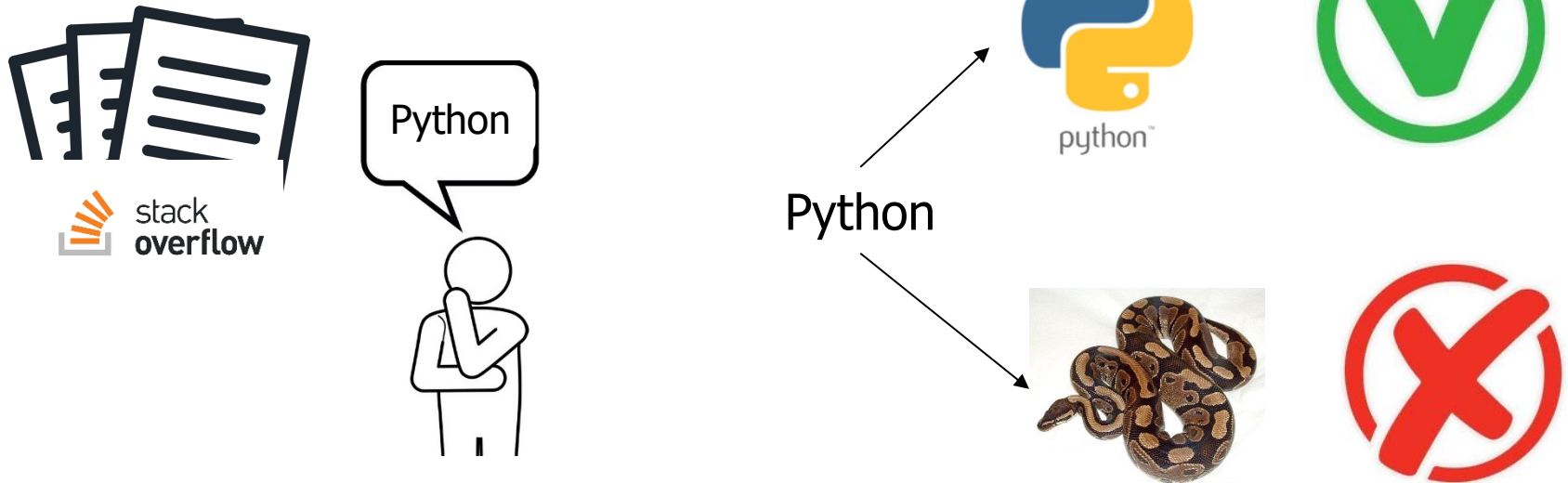
Introduction – Entity Disambiguation

- Entity Disambiguation
  - The process of resolving ambiguous references to entities in unstructured text

Apple

Python
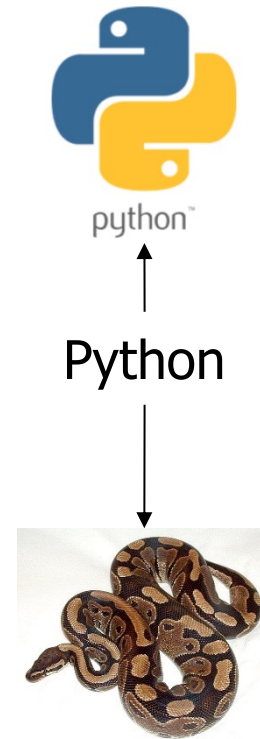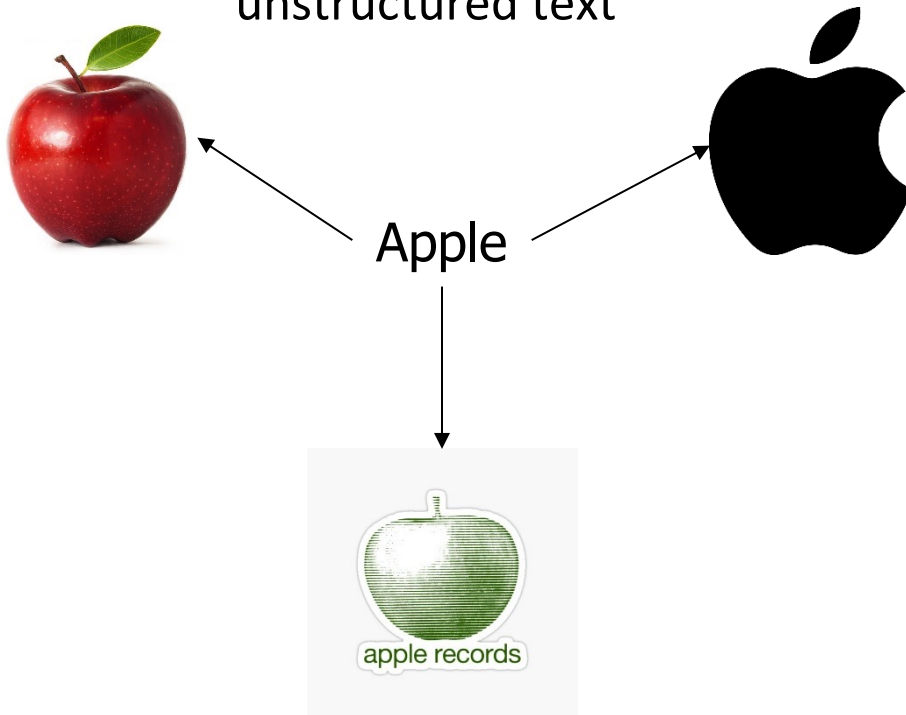
Introduction - Entity Disambiguation

- Knowing the domain can facilitate

Python

Python

- But this is not always the case…

Acute Renal Failure

ARF

Acute Respiratory Failure
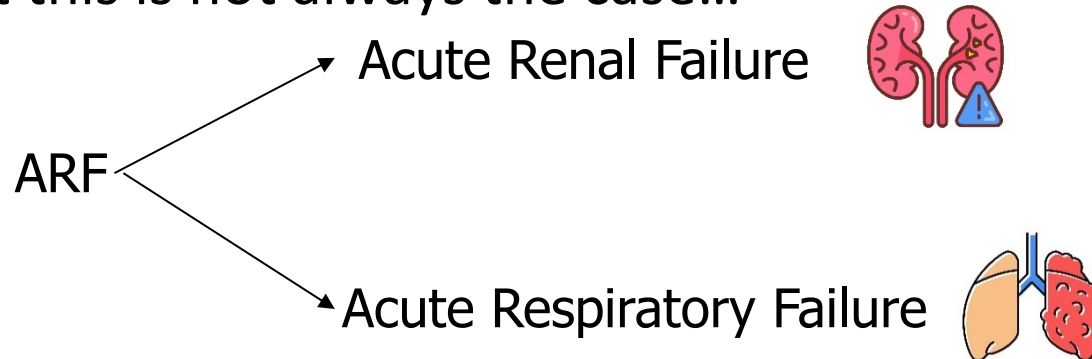
Introduction - Entity Disambiguation

- Entity Disambiguation from Medical Knowledge Bases (KB)
  - Medical KBs are the results of integration from heterogeneous sources

- Discrepancies arise for many reasons
  - Acronyms
    - MRI, CRP, ECG
  - Abbreviations
    - COPD, GERD
  - Typos
    - *Nausia* (instead of *nausea*)
  - Colloquial terms
    - *Renal disorder*, *kidney disease*
      - Both of them are known as *nephrosis*

Clinical resources

Health records

Lab tests
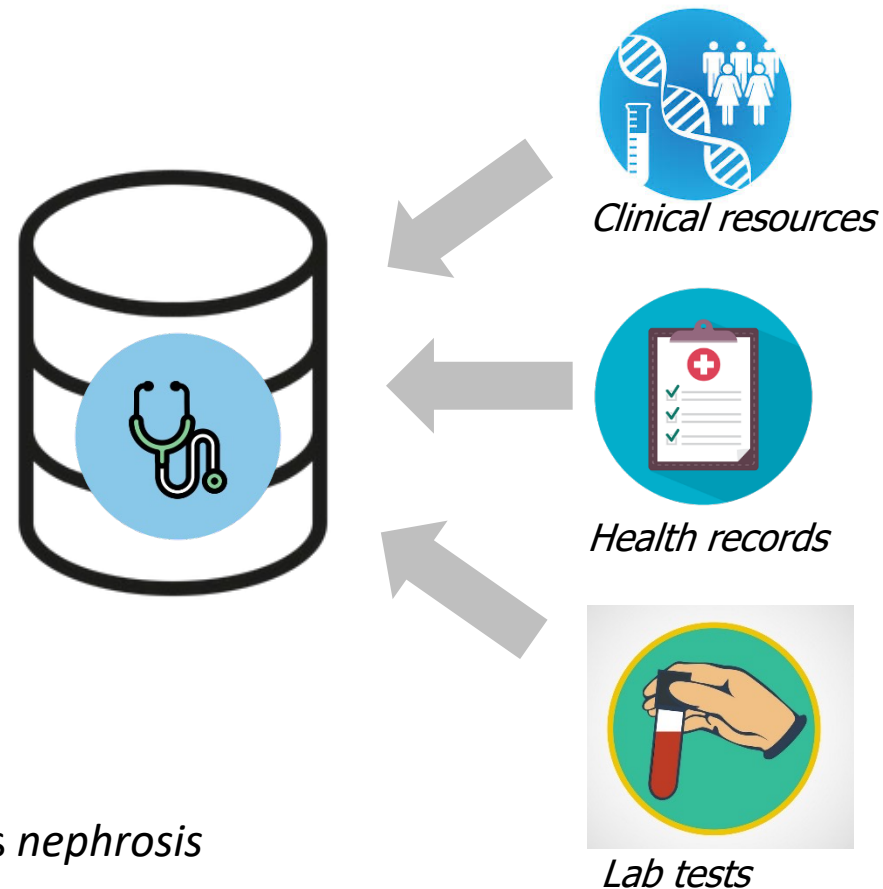
Introduction – Entity Disambiguation, How?

- Rule-based
- Dictionary-based
- Probabilistic models
- Using contextual embeddings
- Machine Learning approaches
- Deep Learning methods
  - Recent studies revealed that they outperform other approaches but
    - Resolve entities based on textual contents (e.g. from surrounding words)
    - They do not fully exploit the structural information in texts snippets and KBs

ED-GNN
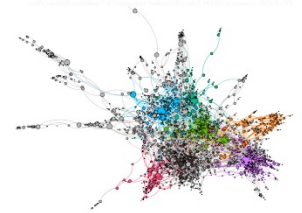
- ED-GNN is a medical Entity Disambiguation system based on Graph Neural Networks (GNN)

- Why GNNs?
  - The foundation of GNNs is a spatial invariant aggregation function that learns how to aggregate rich structural and semantic infromation from each node's neighborhood to generate node embeddings

- How?
  - Entity mentions in a text snippet are likely to share similar or relevant context
  - Entities from a text snippet can be modelled as nodes of a connected graph
  - Entity Disambiguation becomes a binary classification problem using GNNs (i.e. graph matching)

ED-GNN
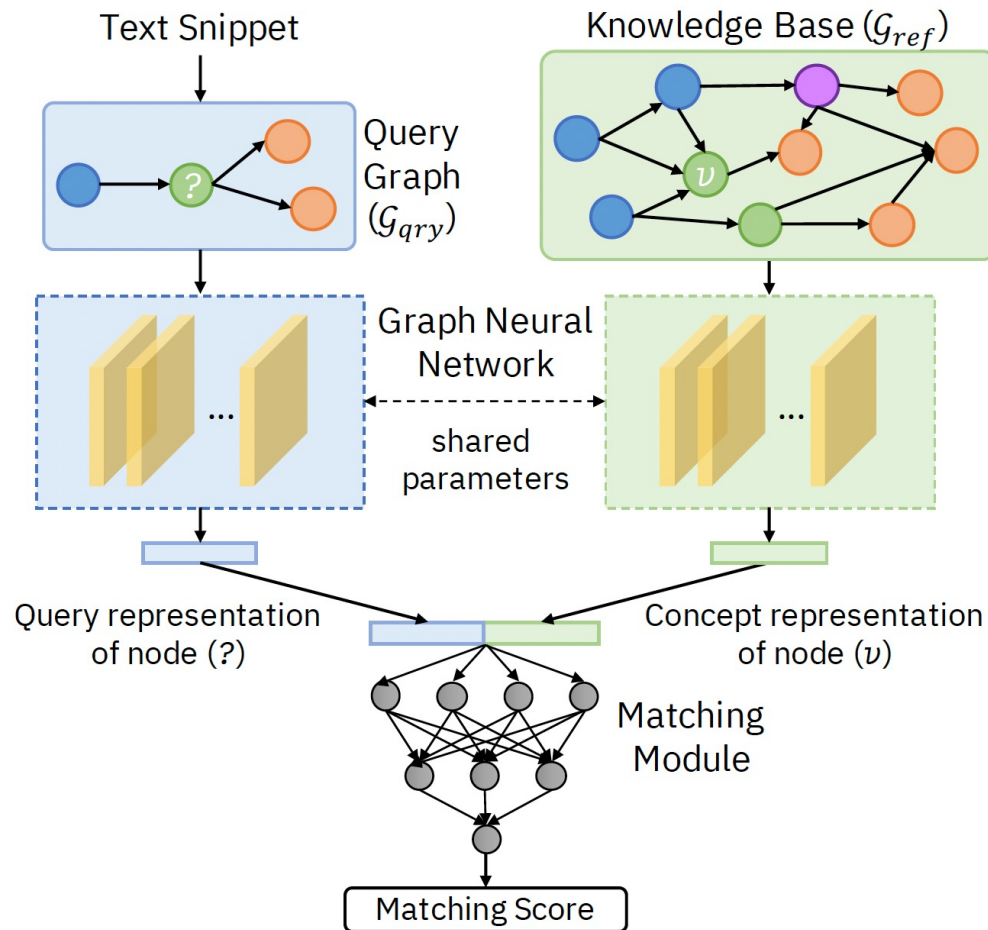
- Represent a medical KB and a text snippet as heterogeneous graphs G_{ref} & G_{qry}

- Medical KBs are curated from medical text corpora (where text snippets come from)
  → G_{ref} & G_{qry} are expected to be similar

- The first step is to identify entities
  → BioBERT which is fine tuned on medical KBs

ED-GNN

- Two major optimizations to fine tune and improve Entity Disambiguation

- OPT1: ED-GNN augments the query graph with domain knowledge from the medical KB
  → Helps ED-GNN focus on the right structural information from the query graph for making the matching decisions

- OPT2: ED-GNN is equipped with negative sampling strategy, which challenges ED-GNN to learn from difficult samples
  → improves the model's dissambiguation capability

ED-GNN – OPT1

- Augment the query graph with domain knowledge from the medical KB
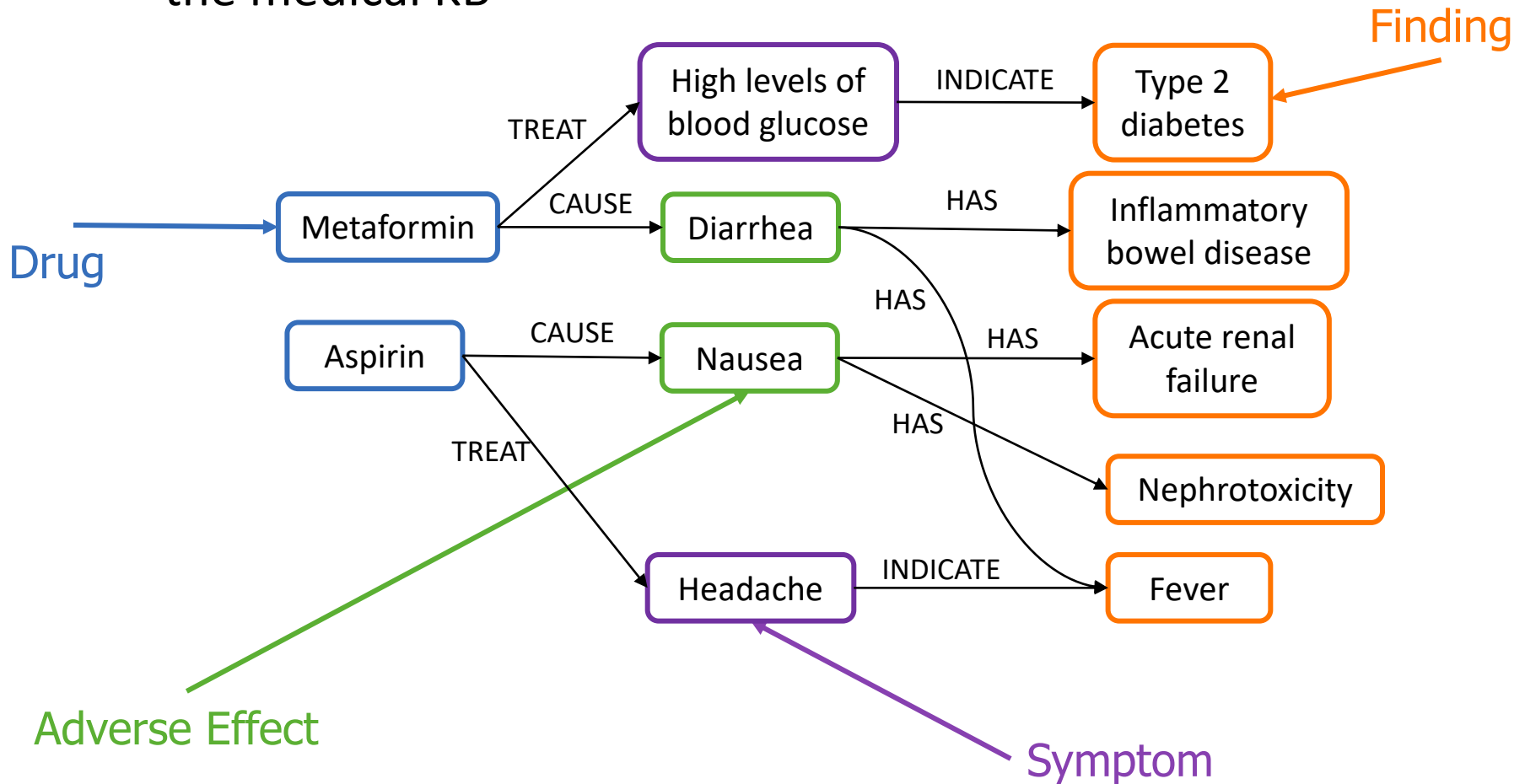
ED-GNN – OPT1

"Aspirin can cause nausea indicating a potential ARF, nephrotoxicity, and proteinouria"

"**Aspirin** can cause nausea indicating a potential **ARF**, **nephrotoxicity**, and **proteinouria**"

"Aspirin can cause nausea indicating a potential ARF, nephrotoxicity, and proteinouria"
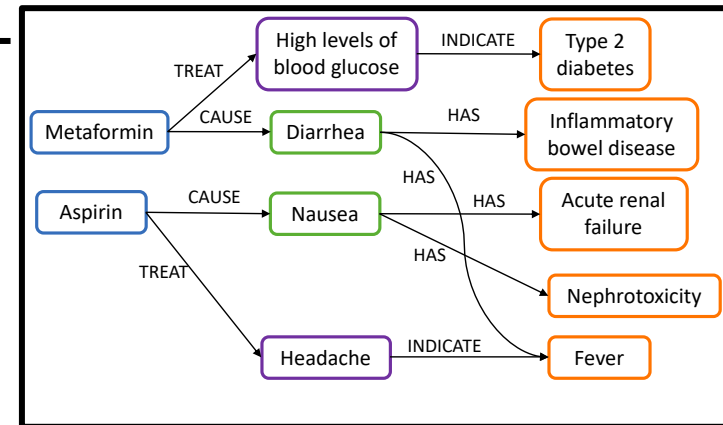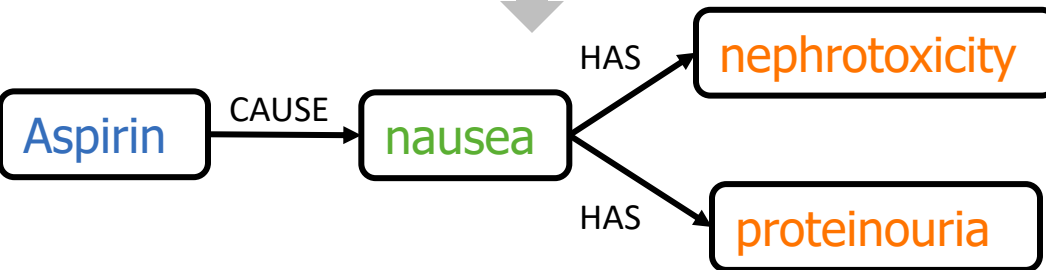
ED-GNN – OPT1
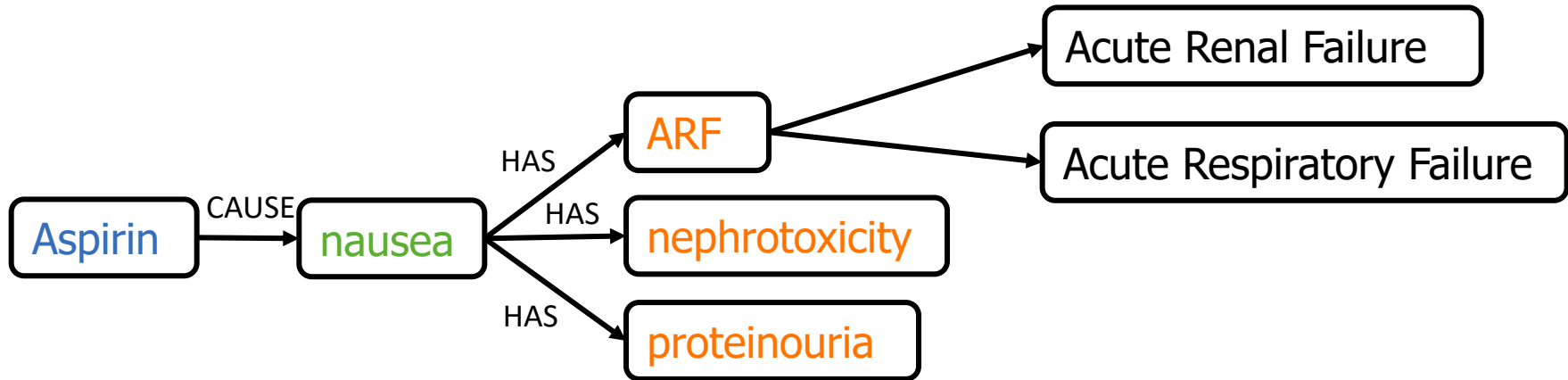
Nephrotoxicity and proteinouria are adverse effects of Aspirin

ARF is amongst the adverse effects of Aspirin

13
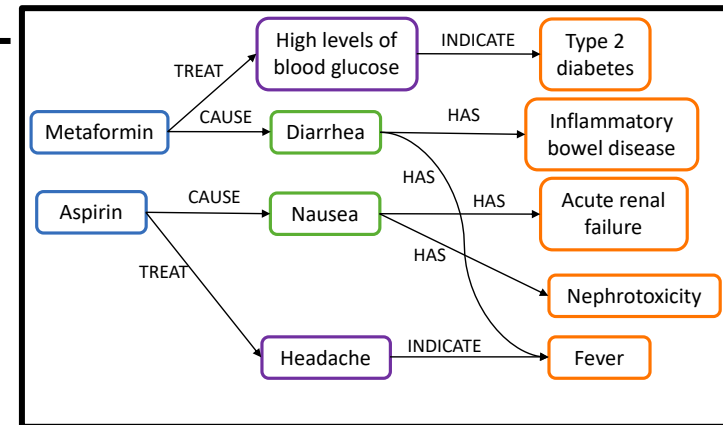
ED-GNN – OPT2

- Negative Sampling is used for improving the training of embedding models
  - Tries to balance the training data by providing a mixture of positive and negative examples
- The most common approach is Random negative sampling
  - Simple but it does not improve the discriminative power of the model
    - E.g. ARF → gastroenteritis
- ED-GNN provides a difficult negative example for evey posititive training example.
  - Negative examples are very close to the positive example entity because of their lexical or structural features
    - Positive example: ARF → acute renal failure
    - Negative example: ARF → chronic renal failure

ED-GNN – Negative Sampling

- Semantic Similarity
  - Positive entity: {MH: malignant hyperpyrexia}
  - Negative Example: {MH: malignant hyperthermia}

- Structural Similarity
  - Negative examples should share many common neighbors with a positive entity
  - To find such examples, Graph Edit Distance is used.

- Gradual Training
  - Easy negative examples (during the first epochs)
  - Gradually provide more difficult examples

- Difficult negative examples are more similar, than random negative examples, thus **forcing the model to learn to disambiguate entities at finer granularity.**

ED-GNN - Implementations

- GraphSAGE, a seminal message-passing GNN
  - Uses messaged-passing techniques and aggregation (nodes aggregate information from their neighboring nodes) to generate graph node embeddings
  - Learns the topological structure of each nodes neighborhood, as well as their distribution

- R-GCN, a relation-aware graph convolutional network
  - Employs also the edges between nodes, distinguishes different neighbors with relation-specific matrices

- MAGNN, a metapath-based GNN
  - Aggregates the node's representation from the metapath-aware neighborhood and the nodes in between.

The equations for computing the embeddings and the state in different layers are given in the paper

Evaluation - Datasets

Dataset	Size	Contents	# Nodes	# Edges
MDX	?	Drugs, adverse effects, indications, findings	35,028	74,621
MIMIC-II	40K records	Patient-health records, lab tests, diagnoses	22,642	284,542
Bio CDR	1,5K abstracts	Chemical, diseases	1,082	2,857
NCBI	700 abstracts	PubMed abstracts annotated with disease mentions	753	1,845
ShARe	433 notes	Clinical notes annotated with disorder mentions	1,719	12,731

Evaluation - Systems

- **ED-GNN (GraphSAGE)**
  - Models the graph topology through neighbors aggregation
- **ED-GNN (R-GCN)**
  - Handles the relationships between entities in the KB
- **ED-GNN (MAGNN)**
  - Learns the representation of nodes based on their metapath-based neighbourhood

---

- **DeepMatcher**   *Baseline Systems*
  - Deep learning method for entity resolution in tabular data
- **NormCo**
  - Based on a deep coherence model for dicesase entity normalization
- **NCEL**
  - Creates a graph for candidates of mentions and then apply GCN to improve the disambiguation

Evaluation – Gold Standard

- Provided text snippets with only one entity to be dissambiguated

```
{"Text":"A common human skin
             tumour is caused by
             activating mutations",
 "Mentions":[
  {
    "mention":"skin tumour",
    "start_offset":15,
    "end_offset":26,
    "category":"Disease",
    "link_id":"C0037286"
  }
 ]
}
```

Evaluation

Methods	DeepMatcher			NormCo			NCEL		
Datasets	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
MDX	0.656	0.700	0.677	0.687	0.634	0.659	0.673	0.659	0.666
MIMIC-III	0.708	0.567	0.630	0.747	0.692	0.718	0.716	0.624	0.667
NCBI	0.783	0.815	0.799	0.863	0.818	0.840	0.816	0.793	0.804
ShARe	0.694	0.639	0.665	0.726	0.623	0.671	0.753	0.631	0.687
Bio CDR	0.837	0.816	0.826	0.866	0.805	0.834	0.857	0.829	0.843
Methods	ED-GNN (GraphSAGE)			ED-GNN (R-GCN)			ED-GNN (MAGNN)		
Datasets	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
MDX	0.614	0.900	0.730	0.722	0.867	0.788	**0.725**	**0.967**	**0.829**
MIMIC-III	0.786	**0.733**	**0.759**	0.810	0.567	0.667	**0.826**	0.633	0.717
NCBI	**0.924**	**0.856**	**0.889**	0.912	0.823	0.865	0.915	0.861	0.887
ShARe	0.794	0.829	0.811	0.806	0.833	0.819	**0.825**	**0.879**	**0.851**
Bio CDR	0.853	0.845	0.849	**0.896**	**0.867**	**0.881**	0.864	0.853	0.858

- All ED-GNN variants outperform baseline systems on all datasets
- The best ED-GNN offers an average improvement of 7.3% in F1 compared to the other best performing solutions
- All models perform better on NCBI and Bio CDR (simple structure)
- MDX and ShARe has more complex structures, ED-GNN gains is much more significant (15.2 % and 16.4% respectively)
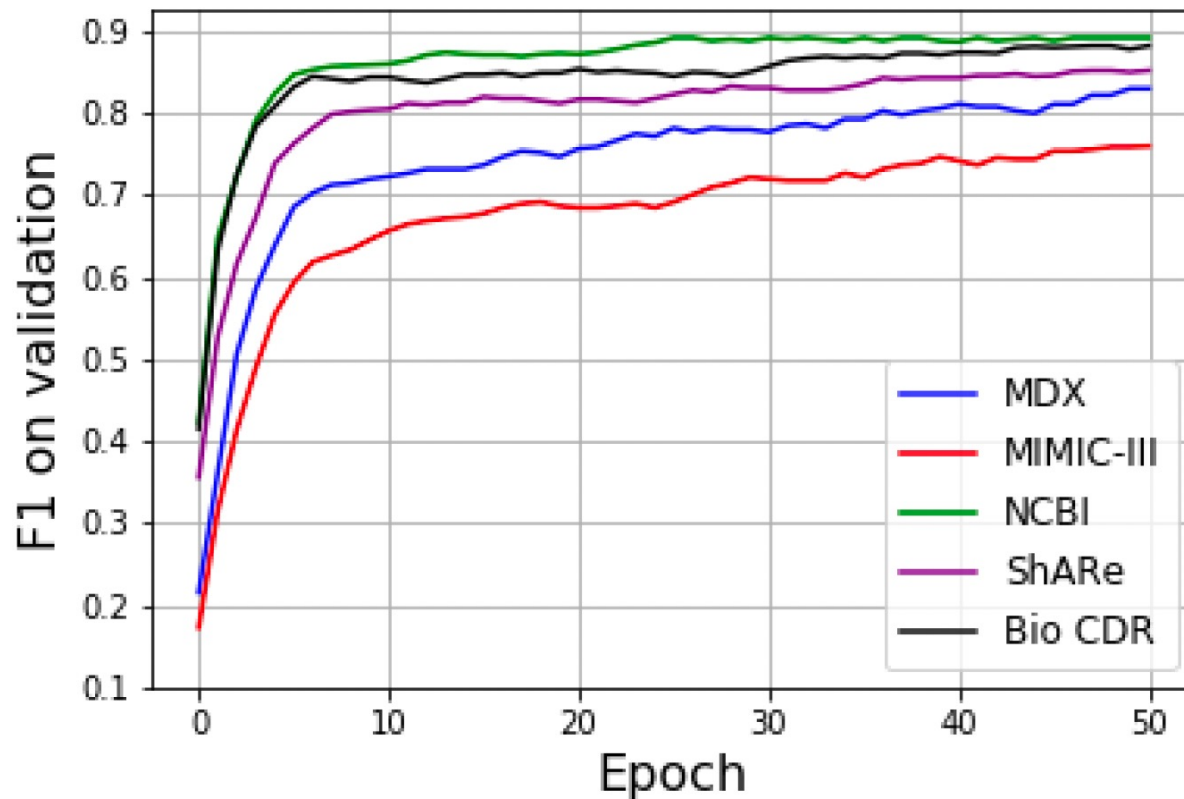
Evaluation

Table 4: Results of two optimization techniques on ED-GNN.

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

- ## ED-GNN (GraphSAGE)
  - Negative sampling improves IR metrics
  - Query graph augmentation does not (GraphSAGE is not a relation-aware GNN)

- ## ED-GNN (R-GCN)
  - Query graph augmention works better (the additional semantic information from the augmented quary graph is more representative when KB is simple)

- ## ED-GNN (MAGNN)
  - Both approaches improve IR metrics
  - Query graph augmentation works better for ShARe (same reason as above)

Evaluation

- Convergence Analysis
  - Using the best performing ED-GNN alternative for each dataset

Evaluation

- F1-Score using different graph layers
  - Using the best performing ED-GNN alternative for each dataset

# 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

- 2 (for NCBI) and 3 (for others) is the ideal
- After 3 layers, the performance declines
  - Distant neighbours introduce much noise leading to more non-isomorphic structures

Evaluation – Error Analysis

- Rate of entity mentions that are not disambiguated correctly by ED-GNN

Error	MDX	MIMIC-III	NCBI	ShARe	Bio CDR
$\mathcal{G}_{qry}$ construction	9.5%	8.7%	1%	3.8%	2.2%
Insufficient structure	4.3%	9.8%	6%	3%	5.2%
Highly similar nodes	8%	4.8%	4%	3%	4.4%

- G_{qry} construction: ambiguous & irrelevant information that lead to incorrect matches
  - *"rash"* can be an instance of Finding or Adverse Effect in MDX
- Lack of structural information from text snippets
  - Graft failure due to FCGS recurrence, in MIMIC-III "Craft failure" is the only neighbor of "FCGS"
- Highly similar nodes in the KB→ ED-GNN not able to learn all possible negative examples through negative sampling

Related Work

- Graph Neural Networks (GNN)
  - Graph Representation Learning has shown to be extremely effective in various domains over graph-structured data
    - GCN, GraphSAGE
  - Heterogeneous Graph Embeddings
    - R-GCN, HAN, HetGCN, MAGNN

- Entity Disambiguation
  - DoSeR, relies on RDF KB embeddings
  - NCEL, creates a graph for candidates of mentions and then apply GCN to improve the disambiguation
  - COM-AID, a composite attentional encode-decode neural network in healthcare
  - NormCo for disease entity normalization

Conclusion

- Entity disambiguation for the medical domain
- ED-GNN leverages state-of-the-art GNNs, further optimized by
  - Augmenting the query graph with domain knowledge from medical KBs
  - Negative sampling to improve the disambiguation capability
- Entity Disambiguation is modeled as a graph matching problem
- Evaluated on top of multiple real-world medical KBs
  - Outperforms state-of-the-art solutions

My Review

- Overall positive comments
- My (negative) comments
  - No measurements and/or discussion about time and computing resources
  - The contents of the medical KB used for the evaluation are (in most of the cases) not available
  - Discussion about adaptations or configuration for applying it for other domains (or open domain)
    - Although they are using BioBERT for entity identification which is fine-tuned on medical KBs

Motivation (0-5)	5
Soundness (0-5)	4
Novelty (0-5)	5
Technical Depth	Diificult (GNNs)
Overal	Accept

References and Links

- Online:
  - https://dl.acm.org/doi/10.1145/3448016.3457328
- Citation
  - Vretinaris, A., Lei, C., Efthymiou, V., Qin, X. and Özcan, F., 2021, June. Medical entity disambiguation using graph neural networks. In *Proceedings of the 2021 international conference on management of data* (pp. 2310-2318).
- Datasets (2 of 5):
  - https://paperswithcode.com/dataset/mimic-iii
  - https://paperswithcode.com/dataset/ncbi-disease-1

Thank You