Medical Entity Disambiguation using GNN

by Alina Vretinaris1∗, Chuan Lei2, Vasilis Efthymiou3∗, Xiao Qin2, Fatma Özcan4

at Sigmod 21/Pods (ACM)

Problem Description:

Entity linking is the task of identifying mentions in documents and mapping them to their correct concept name in knowledge base.Knowledge Base is a huge collection(4.3 B) words in Biomedical Domain.

There are certain issues that follow up while finding an efficient mapping technique and that is resolved in entity linking tasks.

Proposed idea:

**Matching structural similarity utilizing GNN & KB info augmentation.**

Over any basic graph NN , EDGNN adds two optimizations namely:​

1.Query graph –A graph which is augmented from KG based on type similarity and exact matching . For Node Type , Gqry relies on NER task. ​

2.Sampling strategy for Hard negative mining (semantic/Structural dissimilarity score ).

One of the following Graph NN are used in the implemented model :

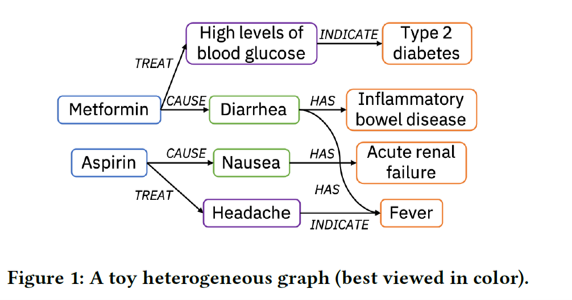
→R-sage – seminal message passing .By using node description etc. learns node embeddings are topological structure of graph.​

→RGCN – relation aware GCN .unlike R-sage ,which considers only nodewise connectivity , it considers also edge labels .In kth layer normalized sum of neighbor nodes is used to update matrix .Different edge type different weights.​

→Mag-NN – Metapath based GNN . For a node . Different Meta Paths on to a node are considered .Uses attention to assign weight to each Metapath.​

Model :

An example of Knowledge graph



Now to construct a similar graph for text snippets , the first NER task is performed.

After that, mentions are assigned as nodes of the query graph and edge types are deduced from Knowledge graph KG.

After that matching algorithm is run which conveys :

Here are two scenario for a node in query graph: It finds its exact match in Gref or it doesn't.​

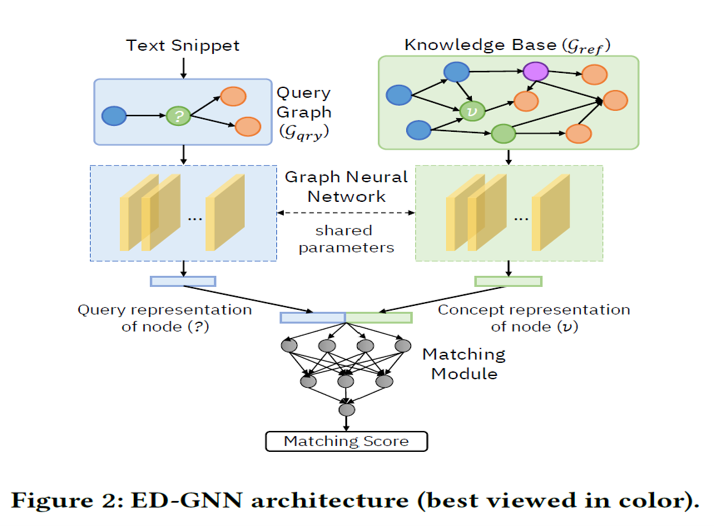
→In the first case exact matching is done to find nodes and edge relationship type and augment to Gqry. ​

→In second case type matching with type set is done first for nodes and then edge type is detected .​

Both GNN should be of same choice.​

Gqry is constructed for some text snippets & Gref is for KB.

Training →



→ A node list, each row contains a node id, its attribute features, and its type.​

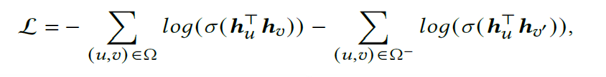
→ An edge list, each row has a source node id (head), a destination node id (tail), and the edge type.​

→pairs are combination of ambiguous mention with its correct and hard negative.​

→The matching module can be a multi-layer perceptron with one hidden layer, a log-bilinear model, or simply a dot product.​

→For hard negative , semantic close or structurally close over 1 hop negatives are poisoned.​

→Back propagation to optimize embeddings.​



Comparison / Results →

Several observations are visible depending on the richness of information in graph structure and denseness of nodes in graph

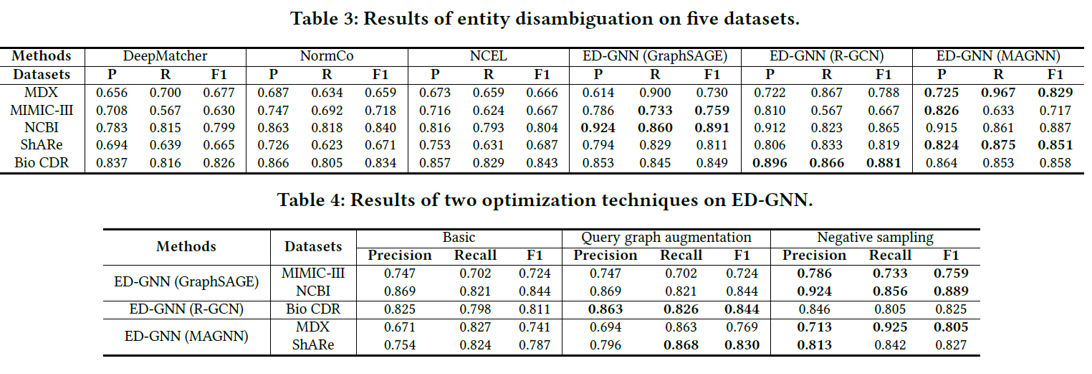
→EDGNN based on GNN such as R-SAGE , RGCN or MAG-NN.​

→16.4% in f1 score.​

→Low complexity & high semantic richness is seen on NCBI dataset so Model performs well . Also dense structural relationships can inhibit the performance. MagNN handles this scenario well.​

→Almost 50% of the errors are due to a lack of graph structural information from text snippets​

​



Bottlenecks:

depending on observations It can be concluded that ::

1. Length of text snippet for which ED is tobe done . It should be large enough to be rich in terms of relationships between nodes.
2. Information richness should be high and complexity of dense regions should be less.
3. Different dataset show different accuracy patterns with GNN choice as can be seen in Table.