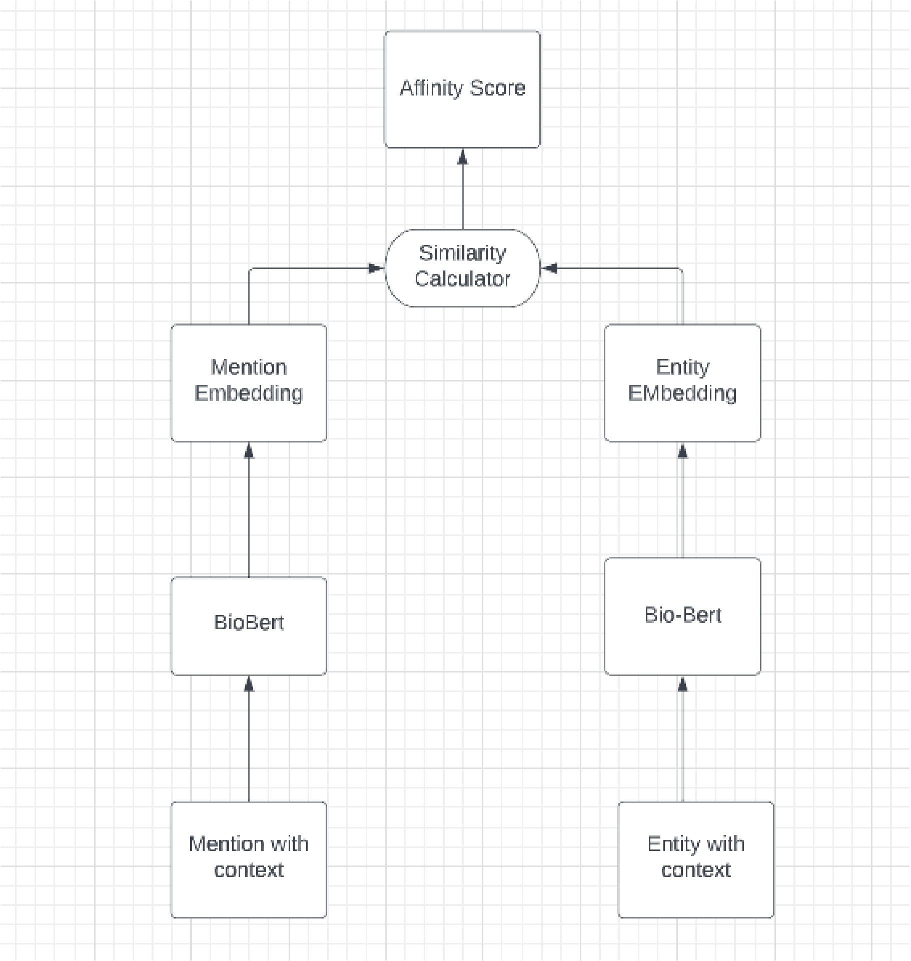
Check

Bio-Medical Entity Linking Summary

The following methods covers high level idea of different Biomedical EL paradigms. In the methods, I have used “*mention*” for mention spans in raw documents and “*entity*” for concept names in KB. Candidate retrieval before reranking is used in all these types so I have not mentioned that separately.

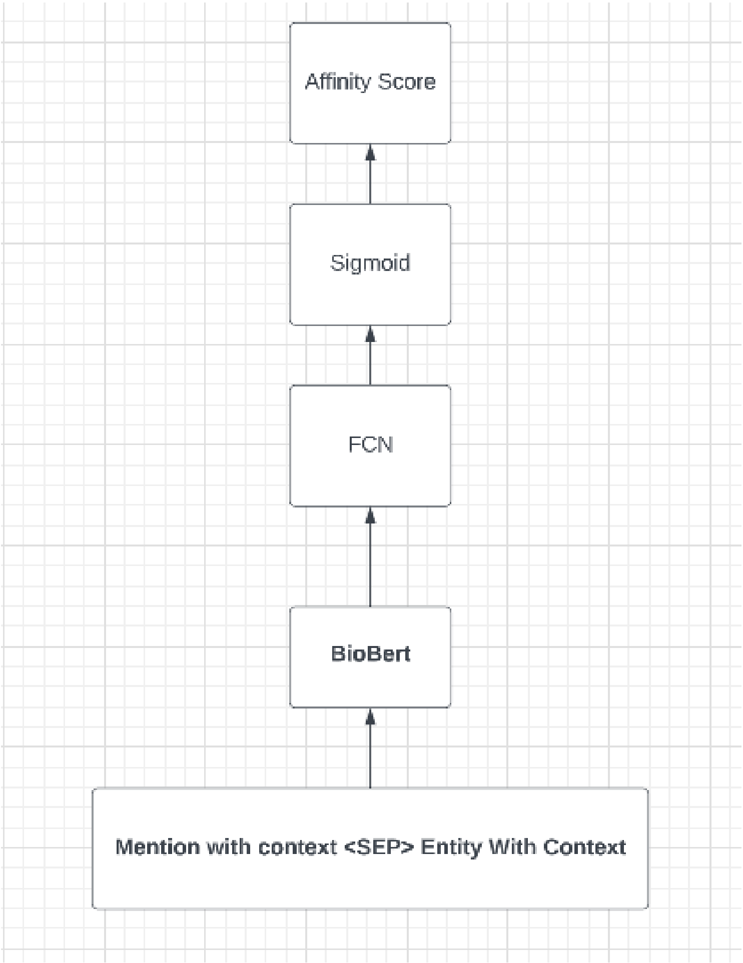
1. **Independent Linking:**
   1.1. **Direct Linking (Bi-Encoder):** Given a mention span with context and entity lexical term, this method independently learn the representation of both using separate encoders and uses a similarity-based method to get the closest entity for a given mention. Hard negatives are taken using the current state of the model at training time. All entity representations are cached at inference.



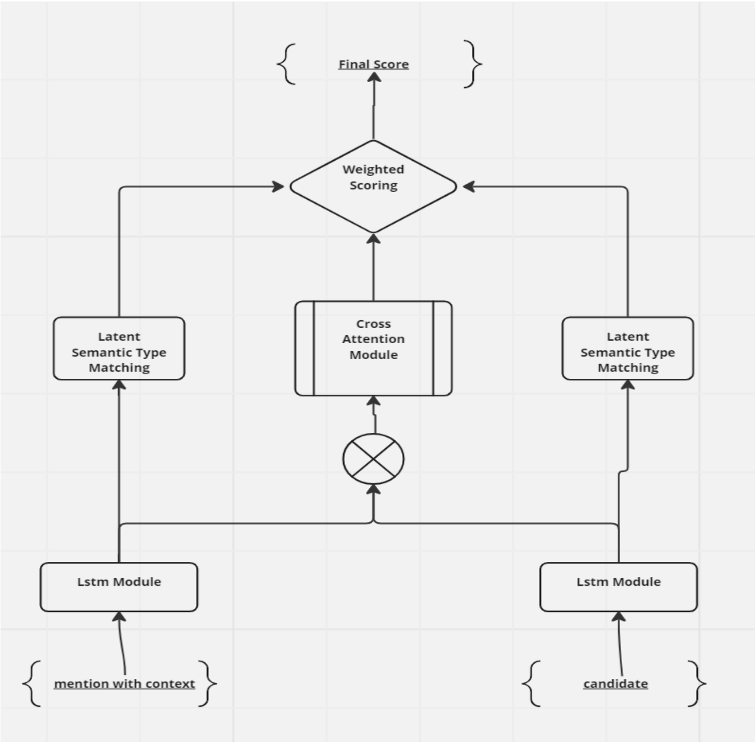
Papers: “Fast and effective EL using Dual encoder” , “Biomedical Entity Linking using RCNN”

In this paper – random & hard negative sampling is being done during training. At inference(testing) time candidate retrieval is not reqd. all entities are precomputed & cached and max. score candidate is selected among them. Dataset – BC5DR & Med mention. Normalized setting is when candidate fetch is done at test time.

1.2. **Direct Linking (Cross-Encoder):** Mention and entity are passed through the same encoder using a special <SEP> token in b/w to get the similarity score between them. It is more accurate but slow.



Paper: “LATTE: Latent Type Modeling for Biomedical Entity Linking”



Here in this paper, other than cross-encoder setup an additional latent type matching module is used and a final score is generated for inference. A score is generated for each entity in candidate set corresponding to a mention. LST module is also trainable using known type classifier module. So, the loss function objective is minimizing score difference btw positive candidate and negative candidate for candidates of a single mention.

Dataset – Med-mention and 3D notes.

2. **Clustering Based:** The idea is to utilize the mention-mention affinity/dissimilarity score along with mention-entity affinity/dissimilarity score to map mentions to entity. Mentions forms cluster with other mentions based on similarity score and links together to their respective entity.

   2.1. **Intra Doc M-M relation:** Works with mentions within the same document. Uses *Cross-Encoder to* get edge weights (**affinity score**) as no. Of mention pairs are less. During training, no clustering is done. Hard negatives (High affinity score) pairs are got by current instant of model during that epoch. Positive pairs are the gold standard pairs. Using this the model is trained. Each cluster will have exactly 1 entity.

2.2. **Inter Doc M-M Relation:** Works with mentions across all documents. During training, a **graph** is created with all mention and entity nodes and edges being **dissimilarity score** between them. Uses *Bi-encoder* as there are a large no of mentions. Extends to ***entity discovery*** if a cluster of mentions does not have an entity present in it. Hard negatives and positives are same as above. Each cluster can have *at most* 1 entity.

Papers: "Clustering Based Inference for BioEL" , "Entity Linking and Discovery via Arborescence-based Supervised Clustering"

In these papers, one uses single cross encoder & another uses dual encoder where a tradeoff of accuracy and time is seen. Clustering algorithm runs only once and objective is to minimize loss function per cluster in base line model. Ground truths are used to train mention-mention model.

In the arborescence-based method directed edges are used to avoid cluttering and the algorithm satisfies that each cluster becomes a disjoint set of minimum spanning arborescence rooted at the entity nodes.

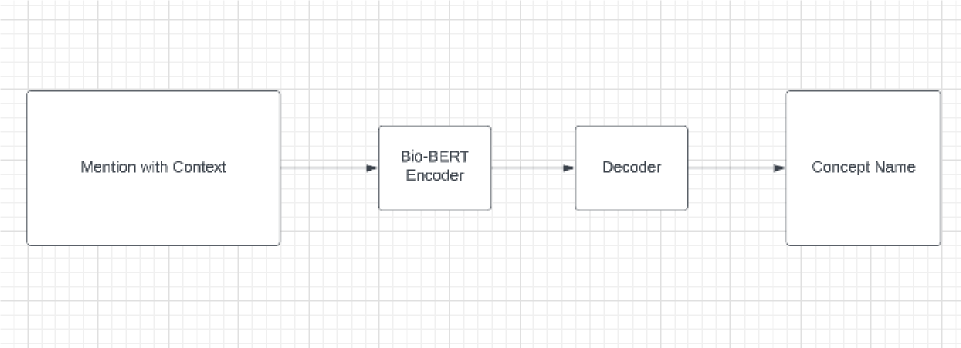
The min-max value is used to assign edge weights to graph unlike in previous paper, and graph prepared is passed to clustering algorithm for sparsification. Loss function objective -

3. Graph Based Approaches:

3.1. Entity-Entity Relation (NeighBERT): A graph is constructed by representing the entity and relationships between entity in the **KB** as nodes and edges, respectively. Each node represents a unique entity concept, such as a disease, drug, or biological process. Each edge represents a relationship between two concepts, such as "is-a", "part-of", or "treats". This helps capture the semantic information of the entities. From this the node embedding is learnt by node2vec or by GCN. Mapping is done by some similarity measure between entity and mention representation.

Papers: "Medical Entity Disambiguation Using Graph Neural Networks", "NeighBERT: Medical Entity Linking Using Relation Induced Dense Retrieval" , "Disease Normalization with Graph Embeddings" (still reading these papers)

4. **Seq-2-seq Based:** Idea is to feed a phrase (Mention with context) to encoder and decoder tries to auto-regressively generate target concept name token by token with the help of prefix prompt and teacher forcing at decoder side. No need to store any concept entity representation at inference saving heavy RAM usage.



Paper: “Generative BEL using via KB guided pretrain and synonym aware fine tuning”