Fast and Effective Biomedical Entity Linking Using a Dual Encoder ( published april 2021 details -<https://aclanthology.org/2021.louhi-1.4/> )

So to convey what paper is all about , Complete review is in the context of following ques →

* What is the objective / what we are doing
* Some basic terminologies related to problem
* What are the current models/method , how do they tend to solve the problem.
* What are the bottlenecks of this problem.
* What are limitations of current model and challenges for the proposed model.
* Proposed model steps/architecture
* Results and limitations

**Problem Description→**

Entity linking is task of identifying mentions in document and map to their correct concept name in knowledge base.Knowledge Base is a huge collection(4.3 B) words in Bioamedical Domain.

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

KB is a 4.3 b word dataset of biomedical domain.

**Introduction and Overview**→

Task of identifying mentions of biomedical concepts in text documents and mapping them to canonical entities in a target thesaurus.

Existing models are Bio-Bert( trained on Biomedical data) based model which utillizes the single encoder and runs 2 models for each entity/mention in the text to resolve the candidate mapping thus proving to be non efficient but accuracy is good since the contextual learning using attention mechanism is performed.

Proposed model is a BioBert dual encoder based model → resolves multiple mentions mapping in one shot. Also enhanced with entity disambiguation and end to end linking. Purpose of using Bio-Bert is for representation of contextual information of mentions Bio-Bert is a BERT which is trained on bio-medical literature data is used here which generates contextual embedding. Any new word can be expressed by learning its context so outbound entities is not a major problem here

**Terminology →**

**Bottleneck for current SOA models →**

Extraction of concept names in biomedical domain requires expertise , So leads to failure of already present models , basically it leads to

Due to →

Jargons and shorthand representation of words in documents related to medical domain .

Similar Surface concept names are available in KB→ so disambiguition while mapping gets tough . For example, Pseudomonas aeruginosa is a kind of bacteria, while Pseudomonas aeruginosa infection is a disease

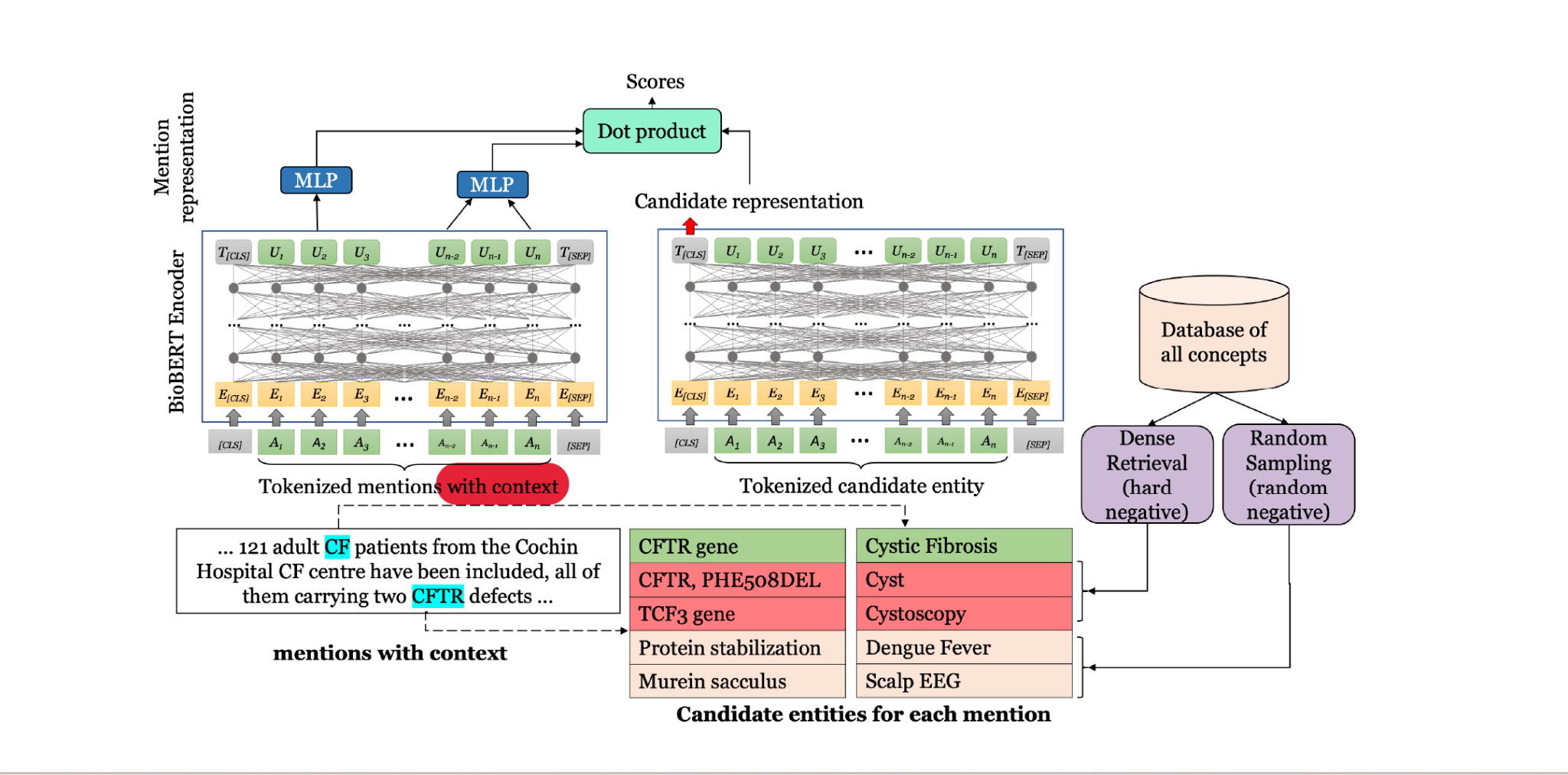
Also → detecting long mention spans and their frequent presence in document ( mention per document ) although not part of linking is challenging task

**Intution & proposed model details →**

So for one shot mapping for the whole document Dual encoders are being used where one is for concept name representation and other is for representation of mentions .

Concept encoder trains and cache the representation of concept names .Also presence of dual encoder facilitates the parallel training part of model and enhances the efficiency at inference time . And instead of reranking , all potential candidates are fetched and cached and at the time of inference dot product (cosine sim) is calculated and once with lowest score from gold candidates are selected.

Figure



**Retrieval of entities step** →

One important step in the process is candidate retrieval .This step fetches all potential candidates which may be the desired mapping .The random negatives are also fetched along with Hard negatives are used as false positives to train models. There are different hard negative mining strategy , The one which utillizes nearest neighbour distance are taken to use to fetch the hard negatives ,or the hard negatives are the one which are more similar to the mention than gold target entity.

Hard negative have important role in the training part since these act as false positives for the KB training. These type of sampling helps the model in optimizing in improving its disambiguition loss .

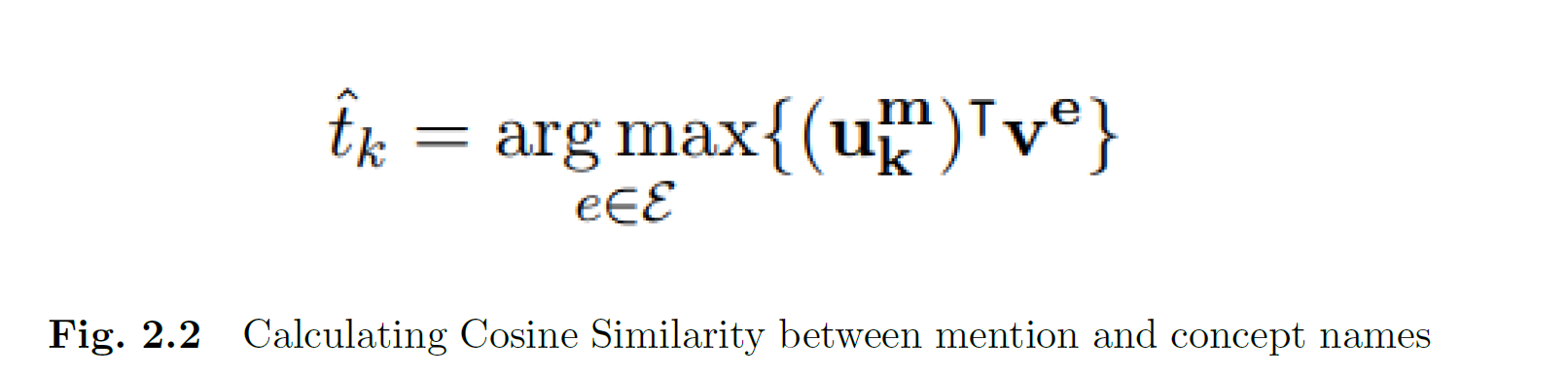
**Encoders and representation** →

U=W[hi; hj]+ b:

**Inference step details**→

In this step scoring is done and prediction step is performed.

for each mention mk are scored using a dot product between the mention representation u(mk) and each candidate representation



**Training** →

The major part of training is to formulate loss function which maximizes score for gold target entity. Binary cross -entropy loss function which marks loss positive for gold entities and negative for negative samples.

Each mention ranging index (i; j) is represented by mean pooling the final layer of the encoder,

.We jointly trained the model by minimizing the sum of mention detection loss and entity disambiguation loss. We use a binary cross-entropy loss for mention detection with the gold mention spans as positive and other candidate mention spans as negative samples. For entity disambiguation, we use the cross-entropy loss to minimize the negative log likelihood of the gold target entity given a gold mention span.

TBC…

**Comparison and results** →

A dual encoder Model which process one mention at a time is baseline here. Actual model outperforms two recent models MedType and SciSpacy with disambiguition related fine tuning .In training 3X , In Inference 20x improvement is observed, without utillizing any semantic info .

This collective dual encoder model outperforms all other models, while being extremely

time efficient during training and inference.

For comparison several models are read 🡪 of which 🡪

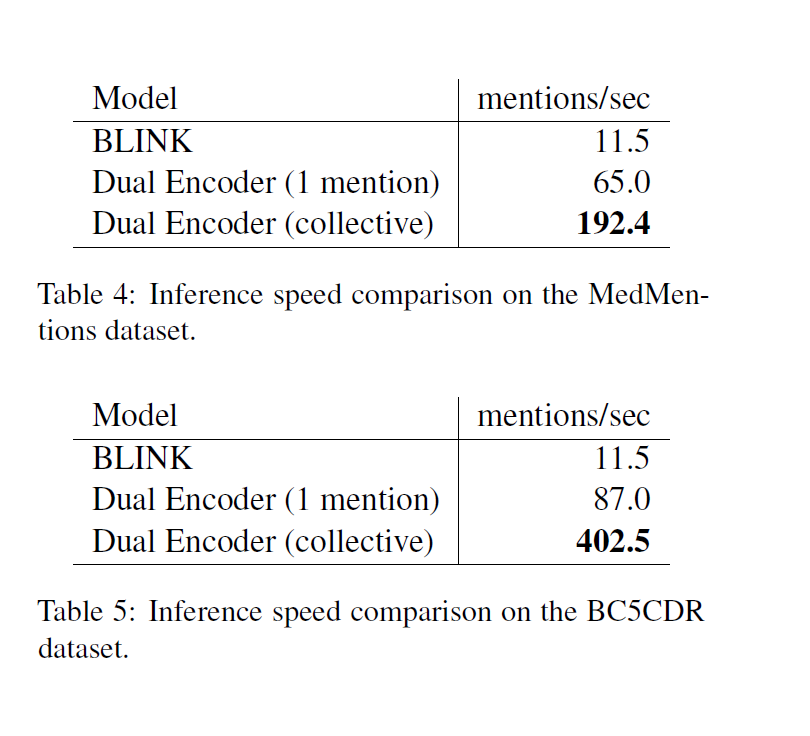
First is LATTE (It adds gold target entity during training)

Cross Encoder model (utillizes cross attention information) (by Lo geswaran)

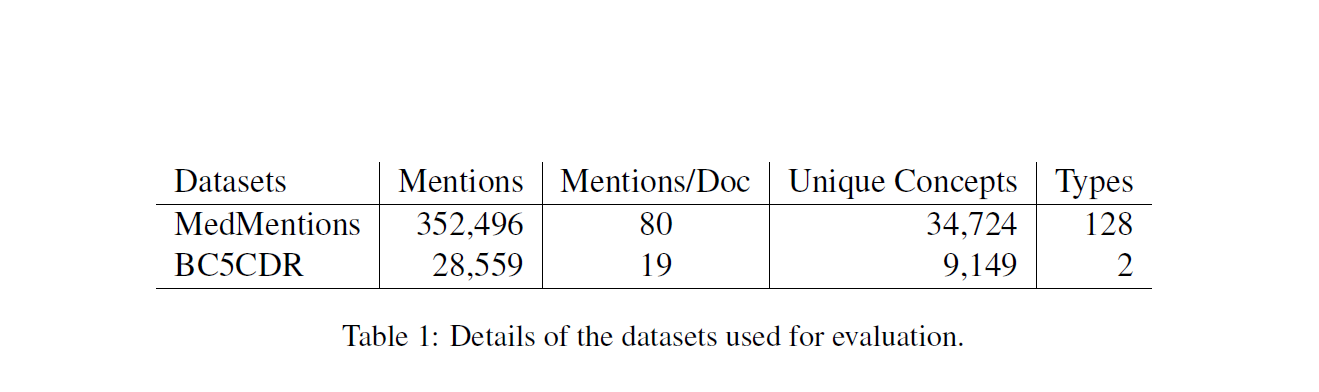
Blink is SOTA model which uses dual encoder for dense candidate retrievel followed by cross encoder for reranking implementation

Our baseline model similar to Gillicks retriever model ,runs for each mention .

Fine tuned model compared with SpiScapy (ngram for end to end linking step ) , MedType ( similar to spacy but uses context information overhead)



**DataSet → MedMention( and BC5CDR whose description can be found on this link**



Bottle Necks and limitation for this problem in general and proposed model → to be updated post referring few more papers.Also training part can be elaborate more.