Clustering-based Inference for Biomedical Entity Linking

from Rico Angell1, Nicholas Monath1, Sunil Mohan2, Nishant Yadav1, and Andrew McCallum1

At

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.

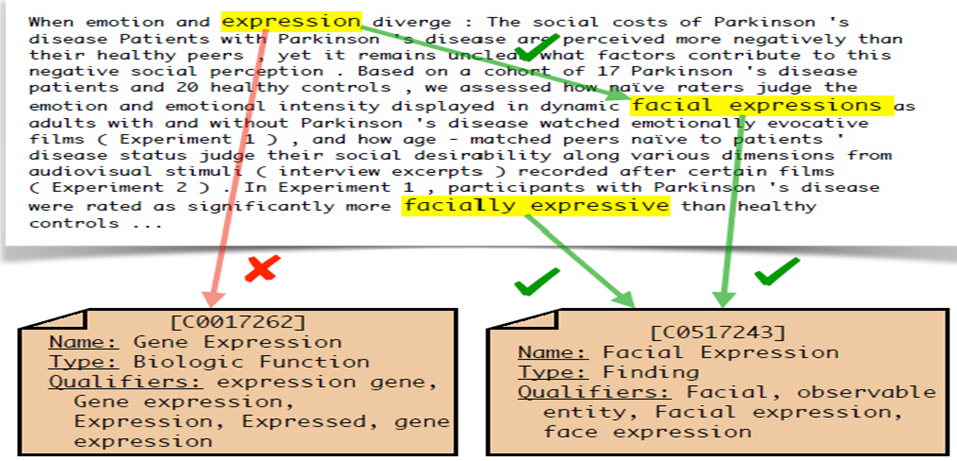
KB is a 4.3 b word dataset of biomedical domain.

Proposed idea:

One technique uncommon in BEL that deserves more attention is clustering.This approach works well when mentions may be difficult to link directly to their ground truth entity, but may have very clear co-reference relationships to other mentions. Most SOTA do not record interaction between intra document mentions , which are similar in nature based on some similarity measure.

This model builds a graph where nodes are union of all of the mentions and entities and edges - have weights denoting the affinities between the endpoints.Each cluster must have one single entity. Mentions are either directly linked or join a cluster for linking to be done.

An example of how clustering helps in linking decision.



Model & Training:

There are two models which are trained in this architecture namely, M-M model & M-E model.

For training MM model , all mention having same ground truth entity are clustered.

Affinity here is defined as the score of similarity between two nodes of a graph.When every cluster contains one entity this process terminates. Initial state is one mention per cluster.There are two trainable models for affinity calc. initialized with embeddings from BioBert.Is a cross encoder taking input as a pair of mention in context and producing a single scalar affinity for every pair.The affinity for a mention pair is computed by concatenating their mention representations(embeddings?) and passing it through a linear layer with a sigmoid activation.

Uses a variant of the maximum spanning tree (MST) supervised single linkage clustering algorithm presented in Yadav et al. (2019).

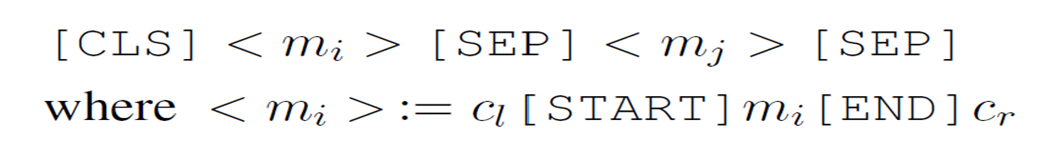
Mention-Mention model runs an algorithm to perform clustering based on scores generated by a trainable model.

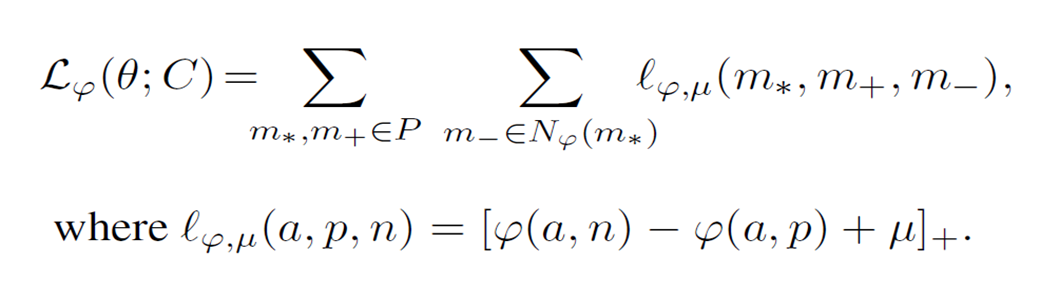
The clustering part is performed at inference time only .

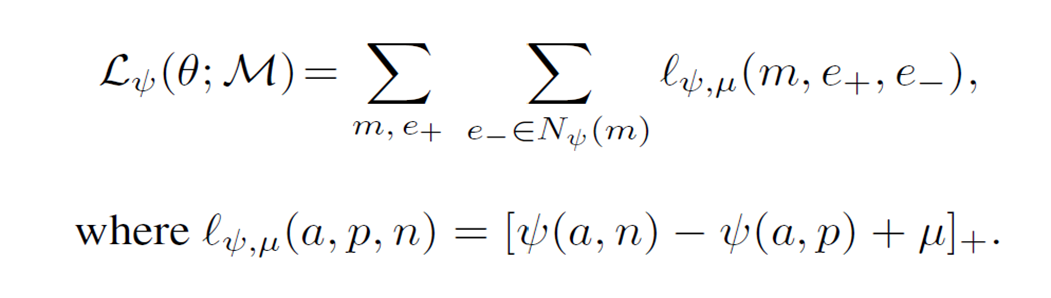
The objective is to minimize the following loss triplet max-margin with margin mu for each cluster C\*.

The objective of loss function is to keep those negatives which are closest from ground truth only i.e. build a closely knit cluster wrt to anchor point.

for each anchor point and positive edge node pair , sum over diff between all negative candidates affinity with anchor and pos with anchor and minimize this sum .(assumption).positive here is ground truth .







N-gram TFIDF model to produce candidates .The candidate generator utilizes the 200k most frequent character-n-grams, and the 200k most frequent words in the names in E to produce sparse vectors for all E of the mentions and entity descriptions.

Bottlenecks:

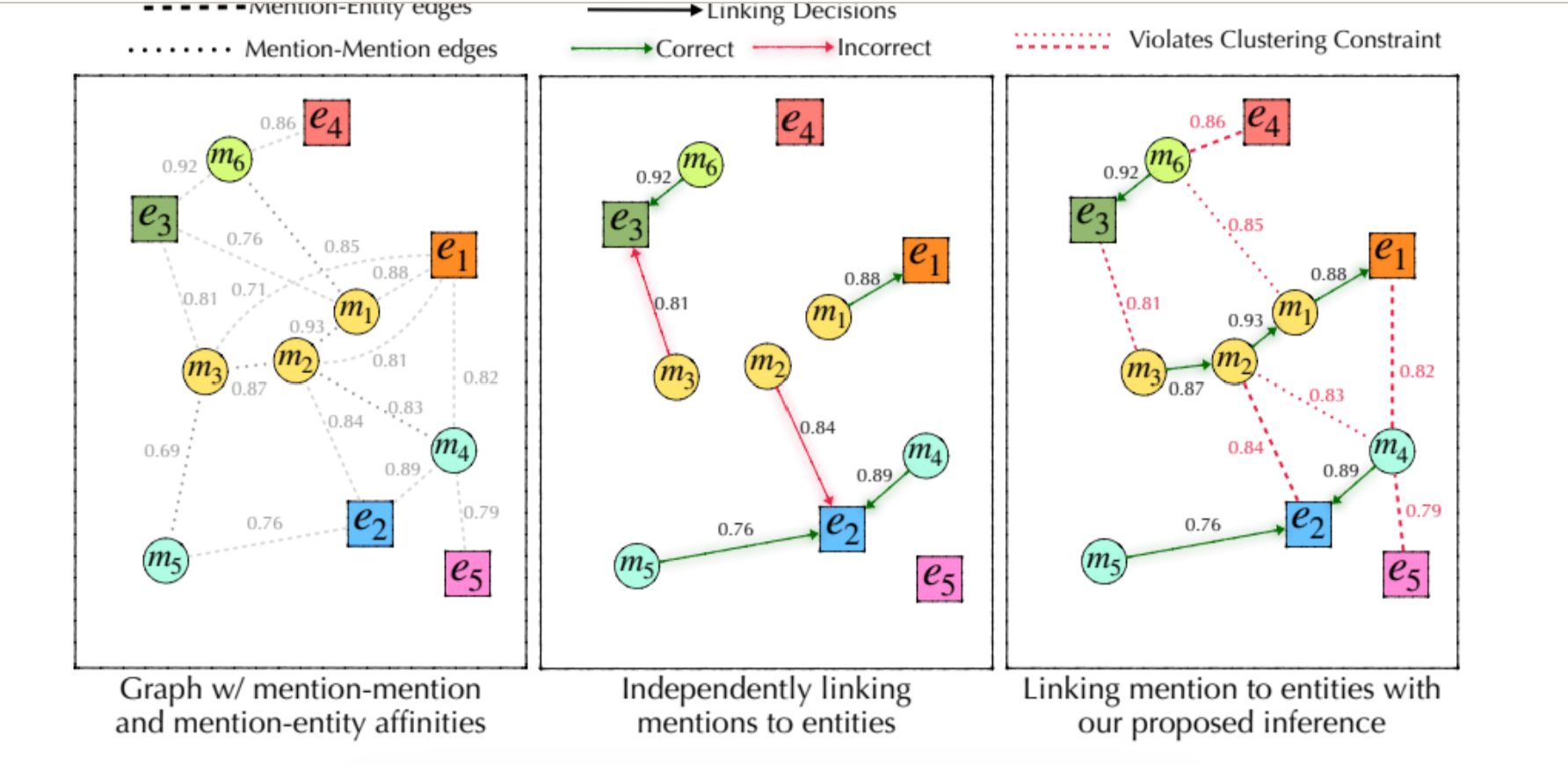
Major part of Training ,Cross encoder & Mention - mention model training .

Although This model shows good results but is time consuming for training .Mainly because of cross encoder used in architecture , which process mention on document level only.

We show through analysis that our approach is better suited to link mentions with ambiguous surface forms and link mentions where the ground truth entity is not in the candidate set.

Mention-mention relationships are also explored in some previous works also. These works use attention in a way to build the context representation of the mentions. However, as mentioned by Logeswaran et al. (2019) is not well suited for zero-shot linking.

Comparison / Results →



An example how this model corrects itself.

