Biomedical Entity Representations with Synonym Marginalization

from - Mujeen Sung Hwisang Jeon Jinhyuk Leey Jaewoo Kangy Korea University fmujeensung,j hs,jinhyuk lee,[kangjg@korea.ac.kr](mailto:kangjg@korea.ac.kr)

At 58th ACL \_2020

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→

In this paper, we focus on learning representations of biomedical entities solely based on the synonyms of entities.By maximize the marginal likelihood

of the synonyms present in top candidates.

Objective to avoid explicit preselection of negative sample to negative samples.

which maximizes the probability of all synonym representations in top candidates.

The candidates are iteratively updated based on our model’s representations

removing the need for an explicit negative sampling from a large number of candidates.

This work is based on marginalizing positivesamples (i.e., synonyms) from iteratively updated candidates and avoids the problem of choosing a

single negative sample.

Model→

We define an input mention m as an entity string in a biomedical corpus. Each input mention has its own CUI c and each CUI has one or more synonyms defined in the dictionary. The set of synonyms

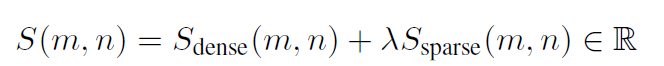
for a CUI is also called as a synset. We denote the union of all synonyms in a dictionary as N = [n1; n2; : : : ]

We first represent each input mention and m each synonym in a dictionary using sparse and n dense representations. A shared encoder is used for both strings.

During training, we iteratively update top candidates and calculate the marginal probability of the synonyms based on their representations. At inference time,we find the nearest synonym by performing MIPS over all synonym representations.

where is a trainable scalar weight for the sparse score. Using , our model learns to balance the importance between the sparse similarity and the dense similarity.

Sparse Entity Representation is done using tf-idf and dense representation is obtained using BioBert.

Final score : 

Training →

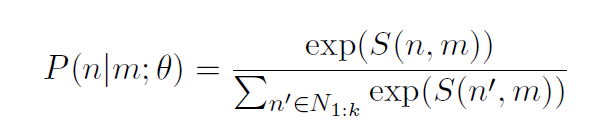
build a pair-wise training dataset. While it is relatively convenient to sample positive pairs using synonyms, sampling negative pairs are trickier than sampling positive pairs as there are a vast number of negative candidates.

First, we compute the sparse scores Sparse and the dense scores Sdense for all n 2 N. Then we retrieve the k 􀀀 bkc highest candidates using Ssparse, which we call as sparse candidates.

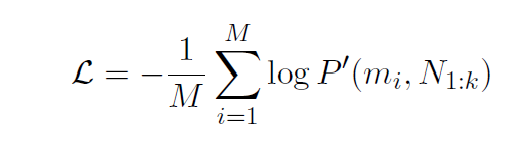
While the sparse candidates for a mention will always be the same as they are based on the static tf-idf representation, the dense candidates change every epoch as our model learns better dense representations.

it increases the chances of retrieving previously unseen positive samples in the top candidates.

Marginalization- we maximize the marginal probability of positive synonyms, which we call as synonym marginalization.



Loss function is given by -



Note that it is computationally cheap to find the nearest neighbors once we pre-compute the dense and sparse representations of all synonyms.

Comparison / Results →

DataSet used - NCBI , Biocreative V CDR , Tac 2017 ADR

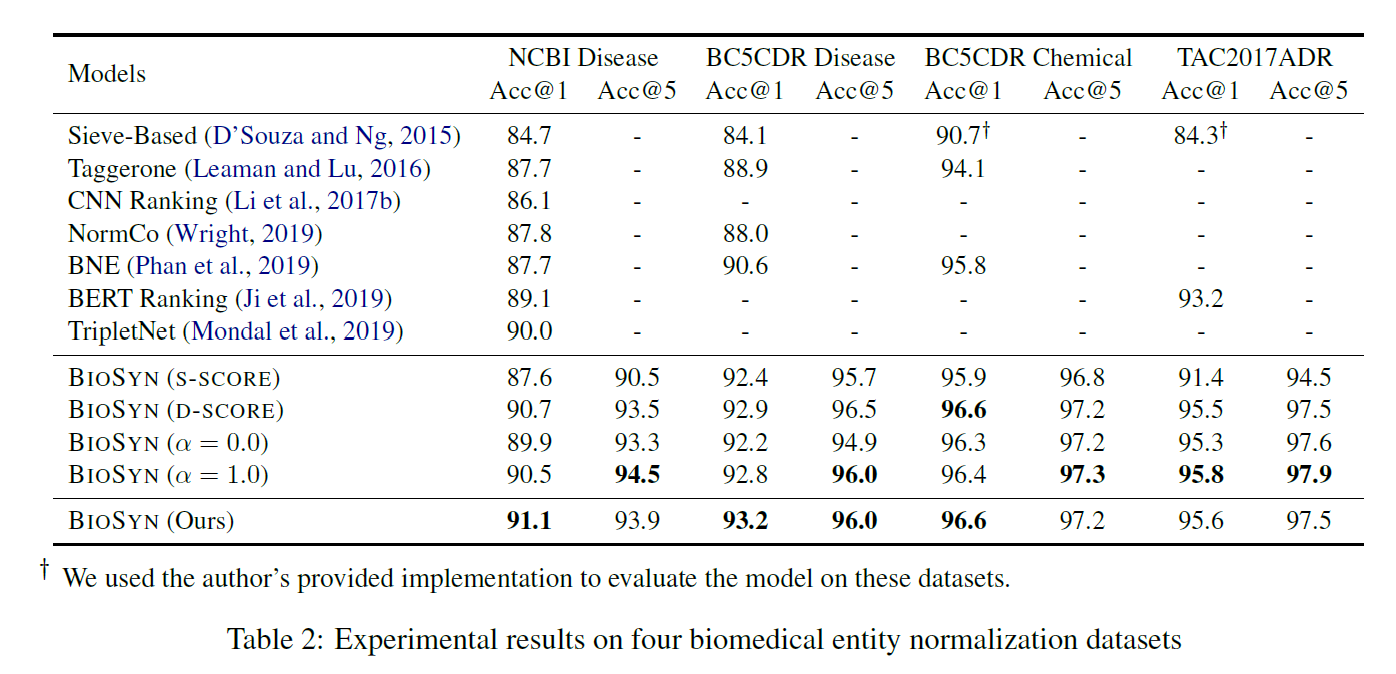
→NCBI Disease Corpus : provides manually annotated

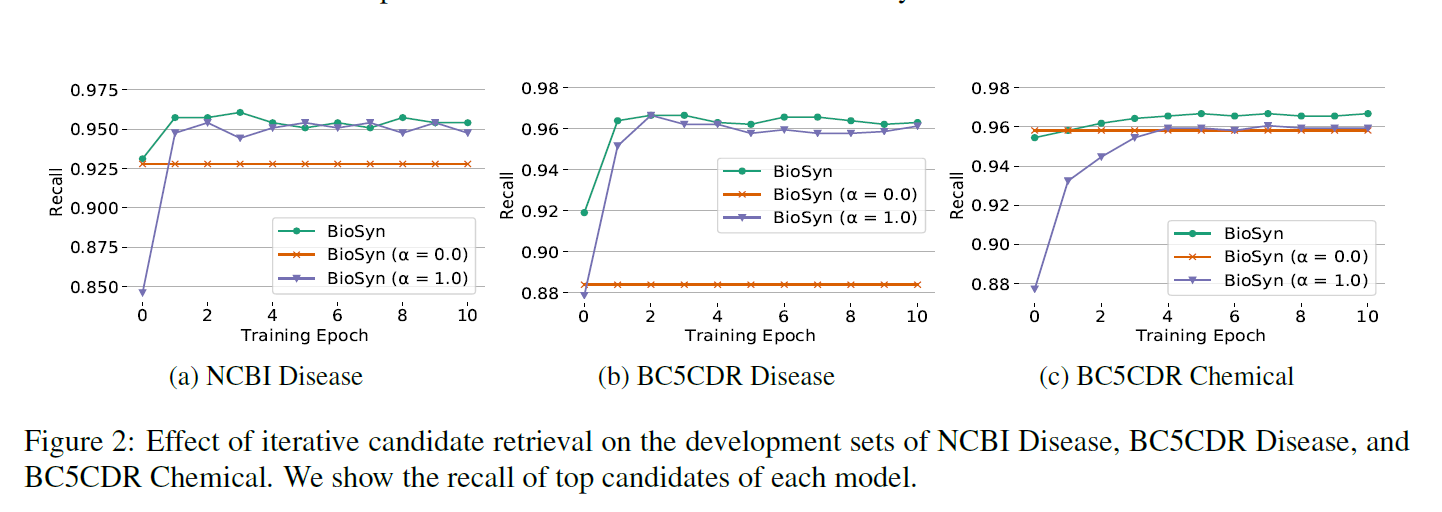
disease mentions in each document with each CUI mapped into the MEDIC dictionary

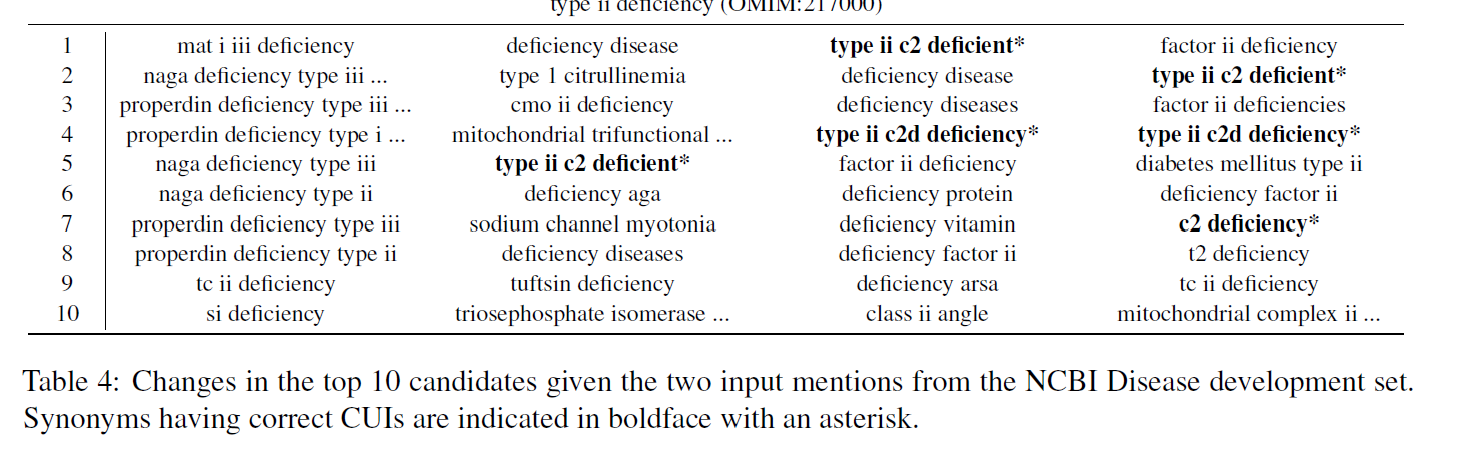
→BioCreative V CDR (Li) : is a challenge for the tasks of chemical-induced disease (CID) relation extraction.It provides disease and chemical type entities.

→TAC2017ADR is a challenge whose purpose of the task is to extract information on adverse reactions found in structured product labels. It provides manually annotated mentions of adverse reactions that are

mapped into the MedDRA dictionary







Bottlenecks→