Could you try to run these two tools on this dataset? <https://raw.githubusercontent.com/obophenotype/cell-ontology/master/cl.obo>

There is a synonym field. We want to match the text between synonym and id. You can do a cross-validation here. Train on part of the synonym, id pairs and test on the remaining.

BiMPM: <https://github.com/zhiguowang/BiMPM>

Densely-Interactive-Inference-Network

https://github.com/YichenGong/Densely-Interactive-Inference-Network

Please find attached the partitioned cell ontology dataset.

cl\_names.tsv contains (idx,name) pairs, where name is a node's name in the ontology, and idx is an integer in [0,#nodes-1].

cl\_train/dev/test.tsv contains (idx, syn) pairs in each row, where syn is the input phrase x, and the name corresponding to idx in cl\_names.tsv is the matched phrase. (idx is the target y).

You can use the method 'load\_dataset' in data.py to load them. You may need to use the BERT tokenizer (in tokenization.py) and BioBERT's vocabulary (

vocab\_cased\_pubmed\_pmc\_30k.txt

) for tokenization. Please adjust the directory path to load them in 'load\_dataset'

For evaluation, the metrics are acc@1 and acc@5. I didn't write separate scripts for evaluation. You can refer to the method 'test' in train.py, or you can write the codes yourself.

Please contact me if there are any questions. Thanks very much!

I am sorry but I need to describe the data format again in case of misunderstandings.

There are 3 columns in cl\_names.tsv: ID,NAME and NODE\_INDEX, where ID is the entity's id in cell ontology; NAME is a string of the entity's name; NODE\_INDEX is the entity's node idx in the adjacency matrix (I think you do not need to use NODE\_INDEX and the adjacency matrix).

There are 2 columns in cl\_train/dev/test.tsv: LABEL and SYNONYM, where SYNONYM is the phrase to be matched, (You need to compare SYNONYM with all the NMAEs in cl\_names.tsv and find the most similar NAME), and LABEL is the most similar NAME's ID.

You can put tokenization.py and read.py under a directory named 'Ref' in order to load data.py without errors.

Also, please put the downloaded BioBERT in a directory named 'pretrain', and put the tsv files under a directory named 'data'.