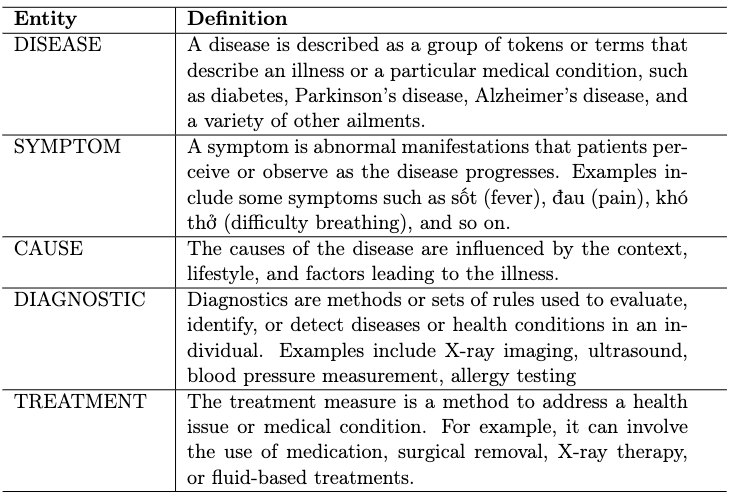
**ViMedNER: A Medical Named Entity Recognition Dataset for Vietnamese**

ViMedNER is a new medical NER dataset that contains 7,749 annotated samples (the sentence level) with five entity types that can be used in common diagnostic and treatment scenarios. Table 1 summarizes the definitions of these five entities. The dataset is annotated by domain experts to ensure the high quality of data annotation. The newly created dataset enriches medical corpora in Vietnamese, a low resource language.

*Table 1: Entity type definitions*



ViMedNER is organized as a txt file, each line is a sample sentence as shown below:

*{*

*"id":11025,*

*"text":"ngoài ra vị trí bị che lấp bởi gan tự nhiên của túi mật tạo điều kiện cho ung thư túi mật phát triển mà không bị phát hiện .",*

*"label":[*

*[*

*74,*

*89,*

*"DISEASE"*

*]*

*],*

*"Comments":[*

*]*

*}*

Where

* “text” is a sample sentence
* “label” is a list of assigned labels. Each entity is assigned a label of the form [start\_character\_position\_index, end\_character\_position\_index, entity name]. As the example above, [74, 89, "DISEASE"], means that the characters 74 to 89, corresponding to the word "ung thư túi mật" is the DISEASE entity.