Clinical Entity Normalization

1. Task Review

Clinical Entity Normalization refers to the linking of different terms or sentences pointing to the same clinical entity to the same term in the standard vocabulary. Our task is to normalize and link such clinical terms from discharge notes that are de-identified and manually annotated with standardized medical vocabularies.

Example: Mapping ambiguous terms like “**cancer**” to “**breast cancer**”, “**lung cancer**”, etc. Or mapping lexically similar terms like “**dilated RA**” and “**dilated RV**”. Or even mapping dissimilar terms “**cerebrovascular accident**” and “**stroke**” which are used in different contexts.

2. Method

- **Proposed Method Description**

  In the previous approach, we used BertForClassification which made the model more dependent on the CUI label, and less on the context. On the basis of feedback, we revised approach to use more context information and no direct CUI information. So, we decided to use Bio-BertNSP as this will make our model dependent on text and will solve the missing CUI issue at time of testing. For NSP tasks, we require two sentences, we take the first sentence as part of the clinical note and the second sentence as definitions obtained from UMLS. We fine-tune Bert with the NSP classification head. While testing, we also extract the score given by the classification head for all candidate CUIs and rank the results.

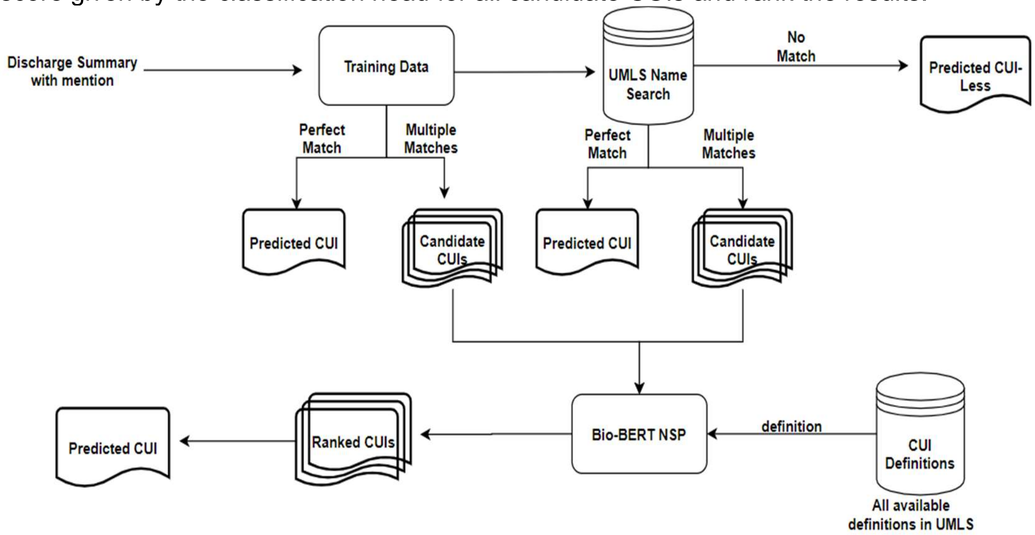

Fig. - Model architecture

- **Innovation**

  Using BertForNextSentencePrediction instead of BertClassification and UMLS database for candidate generation and CUI definitions. Using bio-bert instead of bert-base as it is trained on PubMed data and contains many tokens and references for medical terms.

3. Experimental Setup

- **Baseline Method**

  The Baseline method[1] used UMLS(SNOMED, RxNorm) to predict CUI for a mention. It works in 2 steps, it uses UMLS to find CUIs. If multiple CUIs are generated then BERT with Q&A configuration is used to rank those candidate CUIs. This method generated overall accuracy of 83.56% when using UMLS + BERTQ&A.

- **Dataset Separation**

  Training process uses 50 discharge summaries with a norm file for each which contains the medical mentions and Concept ID. For testing the model, we use gold-standard results provided which have 50 files which have the same format as training data, instead the CUIs are replaced with UNK. In order to fine-tune, we extracted definitions of candidate CUIs from UMLS and this increased our training sample. Initially we had 6780 mentions(iterations) in training data, but after candidate generation from UMLS we have more iterations for NSP. This helps in generating samples that work as negative samples while training the NSP model.

- **Evaluation Metric**
  We used prediction accuracy as a scoring metric for the model. In our case, we aggregated by calculating the mean of prediction score for each candidate and then argmax over it will give the prediction CUI.

4. Result

- **Result of the test set**
  Compared to midterm, where classification of CUIs was done using BertClassification, we used sentences and its definition of candidate CUIs from UMLS in Next Sentence Prediction BERT. So it seems to us that, as we included more context and made BERT not dependent on label/CUI, model performance improved. Taking a more relevant and pre-trained BERT model for this task gives better performance.

Model	Baseline	Mid-term	Final(Bert-Base)	Final(Bio-Bert)
Accuracy	83.56%	57.7%	57.2%	72.4%

- **Error Analysis**
  Contribution of UMLS and training data have a great impact on the performance of the system. The definitions available from UMLS for candidate CUIs helped a lot to produce a good model. Some of the error include:

Category	CUI	Example	Reason
Understanding the second sentence	C0019699	HIV Seropositivity means zvýšená hladina neutralizačních protilátek u jedinců vystavených viru hiv.	The UMLS definitions for candidate CUIs are in different languages.
Incorrect/partial mention	C0085639, C0011923	Falls, Imaging	Not enough data in mention for BERT to learn. Possible mentions for 'Imaging' with more information['Diagnostic Imaging', 'Imaging Techniques', 'Imaging Technology']
Not enough data for learning context	C0428284	"Electrolyte" has 25 candidates, but no definitions in UMLS.	For some CUIs, we only have negative context for NSP, but the actual definition for CUI is missing
preferred names are ambiguous	C4553767 C4761315 C0027498	Nausea, CTCAE 5.0 Nausea Subscale Nausea and vomiting	The prefered terms are ambiguous in themselves. Some include abbreviation, numbers, partial mentions.
Abbreviation, numbers in mention	C0948089, C0027996, C0027853	ACS, VIT . B-3 , neuro w / u	The mentions are themselves abbreviations or contain terms that do not give necessary information.

5. References

1. Unified Medical Language System resources improve sieve-based generation and BERT–based ranking for concept normalization, link
2. Yen-Fu Luo,Weiyi Sun, Anna Rumshiskya, MCN: A comprehensive corpus for medical concept normalization
3. BioBERT: a pre-trained biomedical language representation model for biomedical text mining, https://doi.org/10.1093/bioinformatics/btz682