Contrasting n-gram matching and ClinicalBERT in medical concept normalization

Brian Hur*, Yuxia Wang*
Timothy Baldwin, Karin Verspoor
The University of Melbourne, Melbourne, Victoria, Australia
*contributed equally as first authors

Background:

- Named Entity Normalization (NEN) involves linking named entities to concepts in standardized ontologies, allowing for better generalization across contexts.
- 2019 National NLP Clinical Challenges (N2C2) task 3 focused on Normalization of Medical Concepts in Clinical Narratives
- This involved mapping named entities from the narratives to the correct Unified Medical Language System (UMLS) concept identifiers

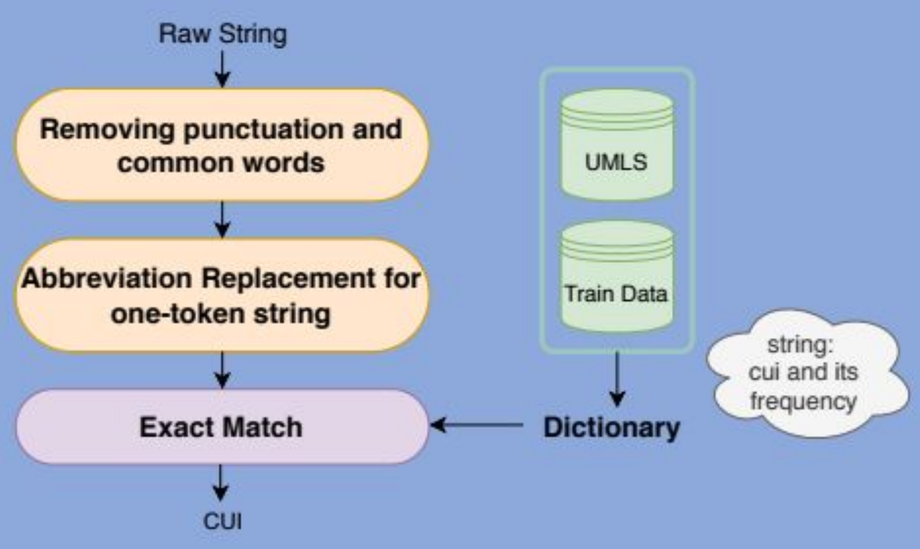
Materials and Methods:

Dataset:

- The N2C2 Task 3 dataset[1] contains **100 discharge summaries** containing **10,919 spans of text** that map to **3,792 unique concepts** from SNOMED-CT and RxNorm
- UMLS Libraries used consisted of SNOMED and RxNorm which consisted of over 1 million concepts

Pre-processing steps:

- Dictionary of SNOMED-CT and RxNorm concept IDs created
- Utilized abbreviation matching from dictionary created from list of abbreviations from NSW Health[2] for text spans which only contain a single token. Example: “PFT” span replaced with “pulmonary function test”
- Perform exact matching of spans to concepts after removing common words and case folding the spans

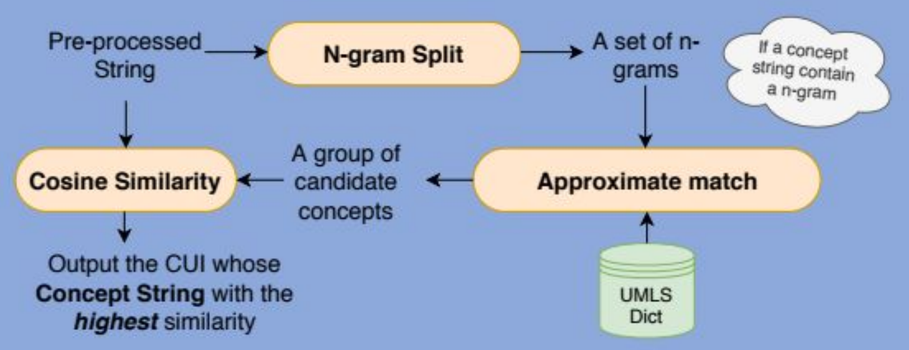


Two separate methods developed for the concept strings that are not matched exactly:

- Handcrafted dictionary-based method leveraging UMLS lexical resources for concept terms and matching n-grams
- Multi-class classifier using ClinicalBERT’s contextualized embeddings[3] pretrained on clinical documents

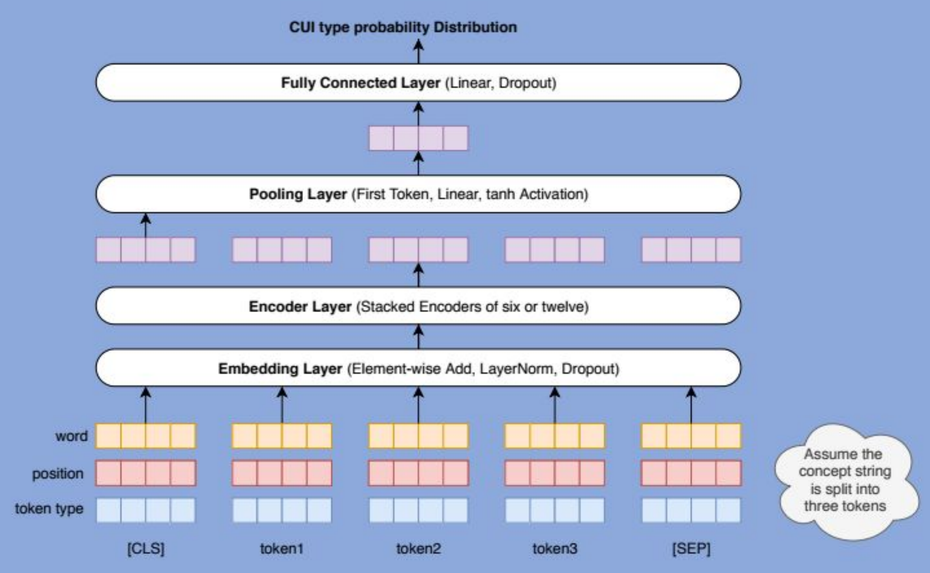
Method #1:

- Generate overlapping word 2- and 3-grams (n-grams of whole tokens). Example: ‘Head computerized tomography scan’ span split into [Head computerized, computerized tomography, tomography scan, Head computerized tomography, computerized tomography scan]
- The sets of tokens are used as query terms for identifying relevant concepts in generated UMLS concept dictionary.
- Cosine similarity between text spans and each concept in candidate list computed based on shared tokens
- Concept ID selected as the concept with highest similarity



Method #2:

- Create target set of unique concepts found in training
- Train multi-class classifier based on ClinicalBERT and original BERT implementation[4]
- Modify classifier to use 2331 unique concept IDs as labels
- Step 1: Use exact match in pre-processing
- Step 2: If no exact match, predict with classifier



Results:

- n-gram based method performed the best with 77.63% accuracy
- ClinicalBERT based method had 77.39% accuracy

Conclusions:

- Two novel methods for N2C2 Task 3 Nomalization proposed; each outperforms baselines for task
- n-gram methods outperformed ClinicalBERT methods by 0.2% accuracy
- Taking the context of concept strings into account is the direction of our future work

References:

1. Luo, Yen-Fu, Weiyi Sun, and Anna Rumshisky. “MCN: A Comprehensive Corpus for Medical Concept Normalization.” *Journal of Biomedical Informatics* 92 (April 1, 2019): 103132. https://doi.org/10.1016/j.jbi.2019.103132.
2. NSW Health - South Eastern Sydney Local Health District. “South Eastern Sydney Local Health District.” Accessed August 6, 2019. https://seslhd.health.nsw.gov.au.
3. Alsentzer, Emily, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. “Publicly Available Clinical BERT Embeddings,” April 6, 2019. https://arxiv.org/abs/1904.03323v3.
4. Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” *ArXiv:1810.04805 [Cs]*, October 10, 2018. http://arxiv.org/abs/1810.04805.