Robust Benchmarking for Machine Learning of Clinical Entity Extraction

Monica Agrawal   MAGRAWAL@MIT.EDU
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA, USA

Chloe O'Connell   COCONELL13@PARTNERS.ORG
Department of Anesthesia, Critical Care and Pain Medicine
Massachusetts General Hospital
Boston, MA, USA

Yasmin Fatemi   YFATEMI@PARTNERS.ORG
Department of Anesthesia, Critical Care and Pain Medicine
Massachusetts General Hospital
Boston, MA, USA

Ariel Levy   ASLEVY@ALUM.MIT.EDU
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA, USA

David Sontag   DSONTAG@CSAIL.MIT.EDU
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA, USA

Editor: Editor's name

Abstract

Clinical studies often require understanding elements of a patient's narrative that exist only in free text clinical notes. To transform notes into structured data for downstream use, these elements are commonly extracted and normalized to medical vocabularies. In this work, we audit the performance of and indicate areas of improvement for state-of-the-art systems. We find that high task accuracies for clinical entity normalization systems on the 2019 n2c2 Shared Task are misleading, and underlying performance is still brittle. Normalization accuracy is high for common concepts (95.3%), but much lower for concepts unseen in training data (69.3%). We demonstrate that current approaches are hindered in part by inconsistencies in medical vocabularies, limitations of existing labeling schemas, and narrow evaluation techniques. We reformulate the annotation framework for clinical entity extraction to factor in these issues to allow for robust end-to-end system benchmarking. We evaluate concordance of annotations from our new framework between two annotators and achieve a Jaccard similarity of 0.73 for entity recognition and an agreement of 0.83 for entity normalization. We propose a path forward to address the demonstrated need for the creation of a reference standard to spur method development in entity recognition and normalization.

1. Introduction

Free text clinical notes contain a rich narrative of a patient's interactions with the healthcare system. However, the strengths and weaknesses of clinical notes lie in the diversity of natural language. Due to the flexibility of documentation, notes contain information omitted from the structured clinical record, but the language itself is often hard to parse into a consistent structured format. For example, `Cold` can refer to the temperature, a temperament, the viral infection, or Chronic Obstructive Lung Disease. By matter of preference, a doctor can refer to a patient with a $101°$ F temperature as `running a fever`, `being febrile`, or `having pyrexia`.

In order to transform text into a unified structured format useful for downstream applications, EHR text mining often involves recognizing spans representing concepts (named entity recognition) and mapping these spans to a common vocabulary (named entity normalization/linkage). We will refer to both steps together as clinical entity extraction. An example of this two-step process is shown below in Figure 1. Typically, in medicine, concepts are mapped to terms in the Unified Medical Language System (UMLS), each term denoted by a Concept Unique Identifier (CUI) (Bodenreider, 2004).
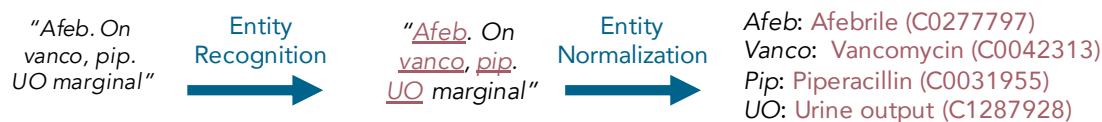

Figure 1: Pipeline of extracting UMLS concepts from text. Typically, this is broken down into two steps as shown: recognition (detecting relevant spans) and normalization (mapping spans to a vocabulary).

Normalized clinical entities can be used as direct input for retrospective analyses and as an additional source of information when addressing confounders. For example, clinical entities can help learn adverse drug events or a health knowledge graph (Jensen et al., 2012; Rotmensch et al., 2017; Nadkarni et al., 2011); or at point-of-care, e.g. as a step in matching patients to relevant clinical trials (Demner-Fushman et al., 2009). Additionally, causal inference and reinforcement learning must both account for unobserved confounders affecting both medical intervention and patient outcome (Gottesman et al., 2019), information often contained in normalized clinical entities. Structured text data could reveal relevant patient history or additional interventions omitted from the structured clinical record. Because normalizing entities condenses high-dimensional raw text into a lower-dimensional space by collapsing synonyms, relevant factors from normalized entities could be plausibly integrated into a model's state space. In a similar vein, Oakden-Rayner et al. (2020) found that machine learning models trained on medical images suffer from hidden stratification, where performance depends on unknown confounders. In particular, the authors found that models for predicting pneumothorax from X-rays were partially picking up on previously placed chest drains. The presence of these chest drains is often noted in the accompanying radiology notes. As before, finding clinical entities associated with abnormally high or low

performance could identify unexpected confounders. Such multimodal sanity checks are particularly important given the high-stakes nature of machine learning for healthcare.

Despite the importance of clinical entity extraction for a range of tasks, it still remains largely an open problem, and clinical research often has to resort to manual chart review for high-fidelity data, which is tedious and difficult to scale. While there has been a significant, concerted effort from the clinical natural language processing (NLP) community to bridge this gap, performance falters due to (i) the lack of sufficient annotated training data, (ii) the huge space of clinical entities to map to, and (iii) known issues with medical vocabularies.

Contributions

In this work, we investigate, quantify performance of, and indicate areas of improvements for the current state of normalization by analyzing performance on the Medical Concept Normalization (MCN) Corpus (Luo et al., 2019). We focus on a subset of CUIs in the UMLS: the SNOMED Clinical Terms, which consist of over 400,000 common clinical concepts; and RxNorm, a comprehensive clinical drug vocabulary (Luo et al., 2019; NIH-NLM, 2015; Nelson et al., 2011). We examine top systems for clinical entity normalization from the 2019 National NLP Clinical Challenges (n2c2) Shared Task as well as two widely used end-to-end open-source systems, cTAKES and MetaMap (Aronson and Lang, 2010; Savova et al., 2010). These analyses reveal holes in performance, underscore the need for future dataset creation and method development in this space, and highlight the need for a new evaluation framework. In their description of the MCN corpus, Luo et al. (2019) recounted how their corpus annotation effort was hindered by several issues including imperfections of UMLS. Therefore, we additionally develop a new annotation framework to address these and other identified issues. We demonstrate how our new schema adjusts to these issues and allows for more flexible evaluation of end-to-end systems.

Generalizable Insights about Machine Learning in the Context of Healthcare

Due to its utility, clinical entity extraction is widely used as an initial step in many machine learning for healthcare pipelines. As the field increasingly turns to multimodal data, clinical text will become a continuously more important component of ML systems. In this work, we show that performance across normalization and end-to-end systems differ significantly on various subsets of the data. Given this observed heterogeneity in performance, it is incredibly important for users of such systems to consider how well they extract concepts relevant to their specific use cases. Unfortunately, a full evaluation of these systems is currently unachievable since a publicly available reference standard for end-to-end benchmarking does not exist. Such reference standards have been crucial to progress in other fields in machine learning, e.g. computer vision through ImageNet (Deng et al., 2010). In this work, we propose recommendations to the clinical NLP community regarding a path forward for the creation of that reference standard. We detail a robust annotation framework that by construction allows for end-to-end evaluation and reduces ambiguity for annotators.

2. Related Work

In the general domain, datasets for named entity recognition and normalization are huge, leading to the success of data-hungry algorithms. The Wiki-Disamb30 dataset, for example, consists of 1.4 million texts (Ferragina and Scaiella, 2010). Techniques also often pull in external information: e.g. population priors, term-term co-occurrence, rich term metadata, and knowledge graphs (Hoffart et al., 2011). Recent successful methods include end-to-end systems and DeepType, which uses a type classifier trained on 800 million tokens (Raiman and Raiman, 2018; Kolitsas et al., 2018). However, these do not translate to the clinical domain because of the dearth of rich annotated clinical text. While a dataset of the magnitude seen in the general domain is infeasible in healthcare, one of our motivations for the creation of a larger dataset stems from successes in the general domain. Another drawback of the clinical domain is the comparative lack of entity information. UMLS contains a Related Concepts table with entity-entity relationships, including triples such as ('Gematicin', 'is_a', 'antibiotic'). Unfortunately, these documented relationships are unreliable and incomplete (Halper et al., 2011; Chen et al., 2009; Agrawal and Elhanan, 2014; Cimino et al., 2003). As a result, the table cannot be cleanly used to leverage entity-entity co-occurrences or other relationships, and more creative solutions are required.

Due to the importance of clinical entity extraction, there exist several open-sourced clinical entity extraction systems, such as cTAKES, MetaMap, and MedLEE (Savova et al., 2010; Aronson and Lang, 2010; Friedman et al., 2004). These systems are very popular; over the past year, cTAKES has been mentioned in over 300 papers, MetaMap in over 500, and MedLEE in over 150. Many of these papers use these systems as a tool in their analysis. Unfortunately, past work has benchmarked the performance of these open-source systems in limited settings to middling results (Reátegui and Ratté, 2018); for example, Wu et al. (2012) found that recall was below 0.5 for all systems on abbreviations specifically. However, due to the lack of a reference dataset, such benchmark numbers don't exist over a diverse set of terms.

Several challenges in the clinical NLP community have focused on addressing entity recognition and/or normalization (Chapman et al., 2011). The 2010 i2b2/VA challenge released a dataset of 871 discharge summaries for recognition (but not normalization) of problems, tests, and treatments (Uzuner et al., 2011). The 2012 i2b2 challenge released a similar dataset of 310 discharge summaries, that additionally annotated clinical departments, evidentials (e.g. 'complained of'), and occurrences (e.g. 'transfer', 'admission'). State-of-the-art for the clinical entity recognition task leverages contextual embeddings (Si et al., 2019). However, performance there is inconsistent, e.g. with an F1 score of 85.1 for treatment span detection, but only an F1 score of 66.3 for occurence detection.

Some clinical datasets also include span normalization. ShARe/Clef eHealth 2013 Task 1, SemEval-2014 Task 7, and SemEval-2015 Task 14 provide normalization to CUIs, but only for disease/disorder concepts (Elhadad et al., 2015; Pradhan et al., 2015; Suominen et al., 2013). Recently, Luo et al. (2019) derived the Medical Concept Normalization Corpus (MCN) from the spans of 100 of the discharge summaries from the 2010 i2b2/VA challenge. They normalized over 13,600 entity spans (problems, tests, and treatments) to terms from SNOMED and RxNorm (Luo et al., 2019). Track 3 of the 2019 National NLP Clinical

Challenges (n2c2) Shared Tasks and Workshop utilized the MCN corpus. The challenge yielded 108 submissions, with a top submission accuracy of 85.2%.

In their paper, the creators of the MCN corpus detailed issues they faced in normalizing terms (Luo et al., 2019). For example, if a compound term was not available in their vocabulary, the authors split the term into its constituents instead of labeling it as CUI-less. Since there is no term for `percutaneous drains`, they instead split the term into 'percutaneous (C0522523)' and 'drains (C0180499).' Since they used the spans from the 2010 i2b2 challenge, they recounted that spans were inconsistently annotated and were suboptimal for the normalization task. They also described inherent issues in the SNOMED vocabulary, such as inconsistency, duplication, and missing terms. These issues led them to have to make self-described arbitrary decisions in normalization. In this work, we build on the findings of Luo et al. (2019) to address these issues in an annotation schema.

3. Current Performance of Entity Normalization Systems

3.1. Data

To understand the current state of clinical entity normalization, we analyzed the outcomes of Track 3 of the 2019 National NLP Clinical Challenges (n2c2) Shared Task, run on the MCN dataset (Luo et al., 2019). Participants had to provide a single corresponding CUI for each provided span. In the MCN annotation process, if the original span from i2b2 (e.g. `high-grade fever`) was split up, teams were presented with the split spans (`high-grade` and `fever`), and not the original span. Teams trained on 50 tagged discharge summaries with 50 held out for evaluation. There were 108 total submissions, and here we analyze the outputs from the top 10 teams. Top teams used a variety of techniques ranging from classical NLP and information retrieval techniques to modern neural approaches, like contextual embeddings. To place performance in context of well-known systems, we compare to output from the Default Clinical Pipeline from cTAKES version 4.0.0 and the 2018 release of MetaMap (Savova et al., 2010; Aronson and Lang, 2010).

3.2. Analysis Methodology

We evaluate performance of the top 10 performing systems on the test data using 3 metrics: (i) highest accuracy any system achieved, (ii) average accuracy, and (iii) pooled accuracy. We define pooled accuracy as the percentage of spans that were mapped to the correct CUI by any of the top 10 systems. In our analysis, we first preprocessed the text spans to remove possessive pronouns tagged alongside the concept (e.g. 'her diabetes' → 'diabetes'). The training dataset consists of 6684 annotations and 2331 unique CUIs, but the 100 most common CUIs constitute 31.5% of the annotations. Since the data set is heavily right-tailed, we evaluate performance for rarer terms by focusing on the following subsets of data:

- All spans *(All)*
- Multi-word spans *(Multi-word)*
- Spans where the corresponding text is unseen in the training set *(Unseen Text)*
- Spans where the mapped CUI is unseen in the training set *(Unseen CUI)*
- Spans where the text isn't a preferred name or synonym in UMLS *(Not Direct Match)*

- Spans mapped to the 100 most common CUIs *(Top 100 CUI)*
- Spans where the text is more commonly mapped to a different CUI *(Unpopular CUI)*

Unpopular CUI measures how well systems do on rarer CUIs, when there are other plausible options. For instance, in the training set, `pt` appears 17 times: 15 times as 'Prothrombin time (C0033707)', once as 'Physical therapy (C0949766)', and once as 'Posterior tibial pulse (C1720310).' A system that always defaults to 'Prothrombin Time' would have 88% accuracy but no disambiguation power.

Additionally, we investigate how accurately different concept categories, as defined by the UMLS Semantic Types, are normalized. We only consider the categories with at least 50 instances in the test set.

Separately, we analyze the performance of cTAKES and MetaMap. These systems were not trained for this specific task or dataset but are built to provide end-to-end annotations, not just entity normalization. Therefore, to evaluate leniently, we deemed a system correct if it output any span with the correct CUI that overlapped with the test span. If the true span did not have a CUI, any output without the exact span was treated as correct. We didn't factor in any additional spans it might have tagged.

3.3. Analysis Results

While overall normalization accuracy is relatively high, it is much more brittle if we look at performance on subsets of the dataset, as detailed below in Table 1. In particular, if a text span is more commonly mapped to a different CUI, it is very rare that systems recovered the true underlying CUI. Examples everyone missed were for 'Cold Sensation (C0234192)' and for the physiological process of 'Diuresis (C0445398)' where all participants chose the term for 'induced diuresis.' Some fraction of errors came from choosing a concept in the wrong type hierarchy, and many are essentially pedantic, e.g. everyone mapped `tube feeds` to 'tube feeding diet (C0311131)' when the gold label was 'tube feeding of patient (C0041281).'

Table 1: Performance of the top 10 systems on the entity normalization task on the MCN corpus. Accuracy is measured on different subsets of the test data and by the 3 metrics described in Evaluation Methodology

Test Subset	# Examples	Max Acc	Average Acc	Pooled Acc
All	6925	85.2%	81.1%	92.7%
Top 100 CUI	1891	95.3%	91.1%	98.1%
Multi-word	2755	81.9%	74.3%	91.3%
Unseen text	2910	69.5%	60.7%	86.0%
Unseen CUI	2067	69.3%	61.3%	81.9%
Not direct match	1682	64.4%	50.9%	84.0%
Unpopular CUI	114	7.0%	2.3%	14.0%

We find that performance is also variable across the UMLS Semantic Types. For some types, the maximum accuracy was nearly perfect: Pharmacologic Substance (93.9%), Sign or Symptom (94.3%), Quantitative Concept (95.8%). Maximum accuracy was significantly

worse for Therapeutic of Preventive Procedures (77.5%), Medical Devices (78.7%), and Body Substances (73.3%). A full table of these results is provided in Appendix B.

MetaMap and cTAKES underperform compared to competition systems. Over the entire test, accuracy was 59.0% for MetaMap, 47.1% for cTAKES, and 64.6% for both pooled together. This is consistent with past findings in which MetaMap yielded poor performance on the SemEval task for disorder/disease recognition (Riveros et al., 2015). The poor performance of cTAKES can be partially explained by the fact that its Default Clinical Pipeline is not trained to tag adjectives (e.g. severe, mild). However, both systems are primarily disadvantaged by the fact that they are end-to-end systems. For instance, competition participants were provided with the span `breast on the left side` which mapped to the CUI for 'left breast'. In contrast, the end-to-end systems needed to first identify the span themselves, before normalizing them. They only identified the span `breast`, and not the full span, and therefore tagged with the less specific CUI for 'breast.'

4. Annotation Framework

Our analysis clearly found that the performance of clinical normalization systems leaves much room for improvement and that properties of existing corpora make it difficult to evaluate end-to-end normalization. In real pipelines, recognition and normalization are usually not decoupled, and as a result, it's important to be able to evaluate fairly in that setting. Therefore, our goal here was to develop an annotation framework to overcome the limitations of past corpora to allow for a more robust, flexible evaluation.

4.1. Framework Development

The annotation framework was primarily developed by medical residents on our team. We iterated from existing annotation guidelines by annotating notes (from the i2b2/n2c2 dataset and the MIMIC-III Critical Care database from Johnson et al. (2016)), addressing inconsistencies and ambiguities that arose, and noting un-annotated information of particular clinical relevance for downstream systems. Our final framework was grounded both by past issues noted by Luo et al. (2019) and our own observations.

4.2. Amended Framework

We summarize the three largest changes made to existing annotation paradigms changes in Table 2 below. Additionally, we further explain the motivations behind each of the three significant framework shifts. We attach the full annotation framework in Appendix C.

4.2.1. Change 1: Tagging Multiple CUIS due to Term Redundancy

The UMLS vocabularies contain redundant terms. For example, there exist separate terms for *Post extubation acute respiratory failure requiring reintubation* and *Acute respiratory failure requiring reintubation*. However, if a patient requires reintubation, they are by definition extubated. The MCN corpus approached this problem by pre-selecting concepts in certain hierarchies to have priority over others. For example, many measurements of substances exist in both 'Finding' and 'Observable Entity' forms, and the creators of MCN informed annotators to choose the former, if presented with that choice. However, this

Table 2: The three significant changes in our annotation schema, in response to the main limitations we found in current annotation frameworks.

Rule	Justification	Examples
Tag multiple CUIs for a single span if they seem equally valid	We no longer need rules on which CUI to pick. Such rules can be arbitrary and are hard to make universal.	Sputum → Sputum specimen (C0444159) Sputum (C0038056) Stent→ Stent, device (C0038257) Vascular stent (C0183521)
Tag subconcepts if they have a medical meaning	Concept boundaries are subjective and UMLS is inconsistent, but tagging subconcepts removes ambiguity and reintroduces consistency (e.g. for Figure 3)	Severe Asthma Exacerbation: Severe asthma exacerbation → C00038218 Severe → C0205082 Asthma → C0004096 Asthma exacerbation → C0349790
Tag all medical concepts (e.g. anatomical terms, normal findings)	Problems, tests, and treatments are hard to delineate and are insufficient to understanding a patient's trajectory.	NSR → Normal Sinus Rhythm (C0232202) Rheum → Rheumatologist (C033489) PEERL → Pupils Equal and Reacting to Light (C1261138)

could have suboptimal effects on model learning, where, for example, representations for the 'Observable Entity' type are improperly learned. Additionally, such rules are hard to codify over the full domain, so in other cases, an adjudicator over the MCN corpus had to choose one CUI, when the two annotators disagreed. In examining errors from the 2019 n2c2 Shared Task, we did find that many "errors" arose from models essentially choosing synonyms. Therefore, our framework allows annotators to choose more than one concept, where they find relevant. This also allows for a more accurate evaluation, in addition to cleaner signal during model training.

4.2.2. Change 2: Tagging Subspans due to CUI Inconsistencies

Inconsistencies in available UMLS terms complicate generate universal rules about concept boundaries. In the example in Figure 2, we show the concept `Severe asthma exacerbation` has a term in SNOMED whereas `severe COPD exacerbation` does not. Additionally existing public clinical entity corpora only tag the CUI corresponding to the longest span of text. Therefore, in the prior example, in one case, `Severe asthma exacerbation` would be tagged, and in the other only `COPD exacerbation` would be tagged. Then, if our model trained on `COPD exacerbation`, our model would likely only tag `asthma exacerbation`. Under current paradigms, this would count as an incorrect tag. As a result, in our framework, we additionally tag subspans corresponding to valid medical subconcepts, as shown in Table 2. This change would increase the artificially deflated performance numbers from

end-to-end systems in our prior analysis, where e.g., they were counted as incorrect for normalizing `breast` and not `breast on the left side`.
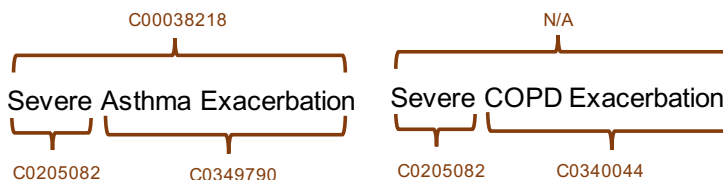

Figure 2: Examples of compound concepts represented inconsistently in SNOMED. While there is a term for 'Severe Asthma Exacerbation', there is no term for 'Severe COPD Exacerbation.' Existing datasets only tag the longest concept available in the vocabulary, but this makes it difficult to annotate concepts consistently.

4.2.3. Change 3: Increasing Clinical Concept Coverage

Finally, we found that problems, tests, and treatments are insufficient to describe clinically relevant information. Problems only cover abnormal findings, but normal findings are necessary to place abnormal findings in context. Further, the definition of a problem is dependent on the history of a patient and the degree of the finding. For example, the s3 heart sound is normal in children or athletes, but can be pathologic in older patients. We face similar issues in delineating treatments, when we may also want to capture a lack of intervention. Consider a note that said *Hypoxemia:...on 3-4L NC [nasal cannula] throughout the day yesterday, weaned to RA [room air] this AM.* `RA [Room air]` isn't a treatment, but it is clinical terminology that is critical to understanding that the patient's oxygen requirement changed.

Finally, outside of issues delineating the three categories, there is other clinically relevant information essential to a patient's narrative, such as anatomical terms and clinical abbreviations. For example, we may return to the example of `pt` which in the MCN dataset could refer to prothrombin time, physical therapy, or posterior tibial pulse. In addition, `pt` is a common abbreviation for patient, and since it is not a problem, test, or treatment, it would never be seen in training; as a result, a downstream end-to-end system would likely face issues. As a result, we believe it is important for datasets to cast a wider net in terms of the clinical terms they tag.

4.3. Comparison to MCN

To provide a point of comparison, we compared a note tagged with our schema versus one tagged with the schema from i2b2/MCN. Under our framework, the annotator identified 170 spans with 123 unique CUIs; in MCN, the note had 72 spans with 63 unique CUIs. The changes in our schema manifested in the following ways:

- `Jaundiced` was tagged both as 'icterus (C0022346)' and 'yellow or jaundiced color (C0474426).'

- The note contained `place percutaneous drains to decompress his biliary tree`. The span in the MCN corpus was `percutaneous drains`, which was split into 'percutaneous (qualifier)' and 'drain (device).' In addition to those, we tagged `place percutaneous drains` as 'percutaneous transhepatic insertion of biliary drain (procedure).'

- Additional annotations included sections of the note (e.g. `Review of Systems`), normal findings (e.g. `stable condition`), and medical descriptions (e.g. `jaundiced`).

5. Validation of Framework

Next, we "validate" our framework to show it can yield consistent annotations and that it had the intended effects on evaluation when applied to real world systems.

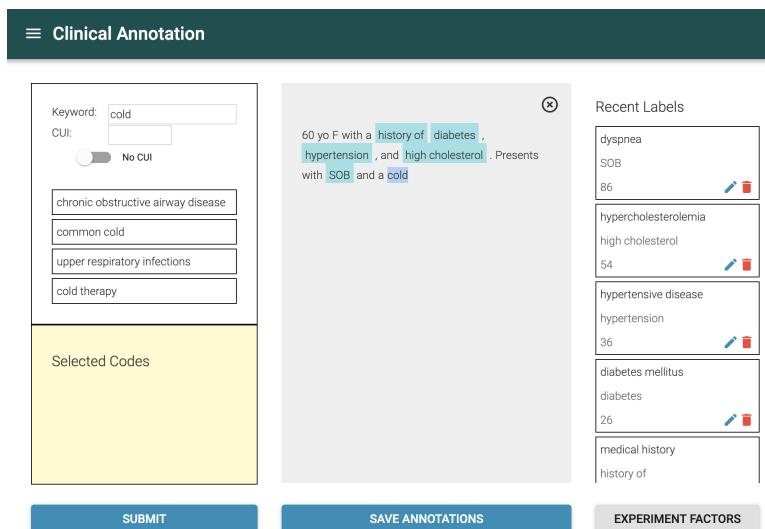### 5.1. Annotating under our Framework

Figure 3: Custom in-house annotation tool, to be open-sourced. Annotators highlight the span they want to tag in the middle panel (pictured: cold), and the panel on the left automatically provides suggestions for CUIs, which can be scrolled through and selected. If the suggestions provided are insufficient, an annotator can use the UMLS browser and directly enter CUIs instead. The right panel shows the labels that were already tagged and at which character locations: the pencil allows for edits and the trash can for deletions.

Two annotators tagged portions of notes, using the MIMIC-III database, so the resulting dataset could be open-sourced (Johnson et al., 2016). They tagged both the History of Present Illness and Assessment and Plan sections of both physician notes and discharge

summaries, since these include the bulk of the narrative and the least auto-generated text. 11 note sections were tagged in all. Annotation was conducted using an in-house annotation tool, which is shown and described in detail in Figure 3.

5.2. Agreement Evaluation

We evaluate concordance between the two annotators by measuring span agreement, and for the concordant spans, CUI agreement. A total of 1571 spans were tagged (1288 by Annotator 1, 1435 by Annotator 2). This corresponds to a Jaccard similarity of 0.73 for unadjudicated span selection. Existing clinical entity recognitions datasets have not released agreement numbers for span selection, so we are unable to contextualize this figure; they merely state that they use two annotators, followed by an adjudicator. On concordant spans, annotators matched on at least one CUI 82.6% of the time (sans adjudication). The MCN dataset reported 67.4% CUI concordance between annotators pre-adjudication. This is not an apples-to-apples comparison, because by additionally tagging shorter subspans of compound concepts, our system benefits from a smaller proportion of compound concepts. However, we also consider this a strength, since it allows compound concepts to be represented in multiple ways to account for inconsistencies in UMLS.

5.3. Evaluation of Real World End-to-End Systems

Finally, we wanted to use the dataset we created to benchmark MetaMap v4.0.0 (rule-based) and the cTAKES Default Clinical Pipeline (hybrid rule and ML model), since they are two common systems used in the machine learning and healthcare community, and past comparisons have only focused on narrow concept types. We examine recall on spans in our dataset, since tools like MetaMap also tag terms like `Date`. We can see results of this analysis in Table 3. For the purpose of the analysis, we considered a union of the annotations created by the annotators, ignoring those spans which did not map to a cUI.

The precision/recall tradeoffs between the two systems are apparent. More interestingly, we find huge discrepancies of accuracies among Semantic Types between the systems. This highlights the utility of benchmarking across a diversity of terms that reflect downstream usage.

Finally, we also examine the 174 spans that contained at least one subspan and were not contained within any larger span. These are exactly the compound concepts that are inconsistently included in medical vocabularies. Of these, cTAKES only recovered the span in 74 cases. However, of the 100 cases with a missed span, in 54 of them, cTAKES did correctly match a CUI for one of the subspans. In existing schemas where only the longest span is tagged, these 54 (31%) cases would have all been counted as incorrect, which vastly misrepresents the true performance of cTAKES. We don't see as drastic a bump with MetaMap, since perhaps in contrast to cTAKES, it is entirely rule-based.

6. Discussion

We have shown that performance of entity normalization is far from robust, and accuracy numbers are hugely bolstered by the existence of a few common, easier-to-map terms. As we have shown, concentrating on certain subsets of the data can lead to vastly different metrics.

Semantic Type	# in Data	MetaMap			cTAKES		
		% Spans Correct	% CUIs Correct	CUI Precision	% Spans Correct	% CUIs Correct	CUI Precision
All types	1271	64%	48%	75%	52%	46%	89%
Sign or Symptom	168	68%	62%	92%	83%	82%	99%
Disease or Syndrome	164	68%	52%	77%	64%	62%	97%
Finding	130	57%	43%	76%	41%	31%	74%
Pharmacologic Substance	113	80%	75%	94%	86%	84%	98%
Organic Chemical	102	79%	77%	98%	88%	88%	100%
Body Part, Organ, or Organ Component	84	40%	22%	55%	72%	66%	92%
Therapeutic or Preventive Procedure	71	58%	30%	51%	38%	24%	63%
Pathologic Function	53	62%	49%	79%	83%	68%	82%
Diagnostic Procedure	43	70%	47%	67%	51%	44%	86%
Temporal Concept	41	54%	46%	86%	0%	0%	N/A

Table 3: Performance of MetaMap and cTAKES on the whole dataset created under our framework, and subdivided by the ten most common Semantic Types. *% Spans Correct* indicates what percent of spans were correctly recognized by the model, *% CUIs correct* indicate what percent of spans were correctly recognized with a correct CUI, and *CUI precision* indicates the percent of correctly identified spans that were normalized correctly.

As a result, performance would be highly variable for downstream applications depending on the entity of interest. This holds true both for the normalization systems from the n2c2 Shared Task and for the widely used MetaMap/cTAKES. Being able to compare systems with a reference dataset is crucial for deciding how to set up a clinical machine learning pipeline. Our annotation framework could be used to create that dataset, since it allows for more nuanced evaluation.

In examining the discrepancies between our annotators, we notice they often arose because of minute differences between concepts, e.g. 'malabsorption' versus 'malabsorption syndrome.' In these cases, note context is insufficient to pick up on subtle differences in definition, which are also generally of little clinical consequence. A system that retrieves either of these should be assessed as correct.

Perhaps most importantly our schema allows for a transition towards end-to-end learning and evaluation. Let us consider an example created from our framework where an annotator tagged `dirty UA` as 'urine screening abnormal (Finding)' and `UA` as 'urinalysis (Procedure).' We can give partial credit to a system that only tags `UA`. In contrast, under current schemas which only tag the longest concept span, normalizing only to 'urinalysis' would be considered incorrect. Not only does this discount the model's true performance, but it also adversely affects training. Further, in this case, the subspan `UA` is of a different

Semantic Type (Procedure) than the whole span (Finding). If a researcher only wanted to extract procedures, under our framework, they won't lose out on those spans where procedures are a subspan of a larger concept. Therefore, our schema allows for flexible training of a subset of concepts of interest.

Limitations

Given that we tag spans in addition to subspans, our annotation framework does greatly increase the number of terms that need to be tagged per document. However, this extra effort is partially offset by the decreased cognitive load of deciding where a span boundary is. Further, by providing suggestions upon highlighting a span, the clinical annotation tool already greatly speeds up the time needed to annotate. Moving forward, we are currently experimenting with having the tool auto-tag unambiguous cases, which the user could simply choose to accept or reject. Since our analysis of n2c2 shows that we can already extract certain terms (e.g. `hypertension`) and Semantic Types with high confidence, this allows the annotator to focus only on the most difficult 20-30% of cases. In manual examination of spans tagged only by a single annotator, differences often arose due to annotation fatigue, rather than misunderstanding of the framework. For example, an annotator would not tag a text span they had previously tagged before in the note. This fatigue could be alleviated by the addition of automatic suggestions.

However, with suggestions and auto-tagging, it is important to understand the sources of bias we could introduce to our resulting dataset. An ongoing area of research is assessing the extent to which the presence, ordering, and presentation of suggestions and auto-tags may bias outcomes. Additionally, we hope to evaluate user interface (UI) techniques to mitigate this bias. There is a need for larger annotated clinical datasets, and smart computational tooling can decrease the time, effort, and cost needed to create them. To that end, we plan to open-source our clinical annotation tool, as well as the annotations generated by this work. See `https://github.com/clinicalml/mimic_annotations`.

We do acknowledge that the annotation guidelines were developed and evaluated using notes written in a critical care context, so that we could open-source the resulting MIMIC data set. Future work remains to evaluate the schema in the context of other clinical settings and note types.

7. Conclusion

Medical notes offer us an incredibly rich view into patient narratives, but they have remained a largely untapped resource for clinical and machine learning research. Clinical studies often resort to manual chart review, which is time-consuming and difficult to scale. While automated concept extraction systems provide a way to exploit clinical text, as we have demonstrated in this work, they cannot yet be reliably deployed.

Developing truly robust and accurate clinical entity recognition and normalization algorithms will require both substantially more labeled training data and automated evaluation metrics that account for the subtlety and ambiguity of the task. We present a new annotation framework and a small-scale annotated dataset for evaluation of end-to-end concept recognition and normalization. The time is ripe for the field to invest in substantially larger

labeled training sets to spur new machine learning approaches for clinical entity recognition and normalization. Future work should consider incorporating the new annotation framework as the basis for such a data set.

Acknowledgments

References

Ankur Agrawal and Gai Elhanan. Contrasting lexical similarity and formal definitions in SNOMED CT: Consistency and implications. *Journal of Biomedical Informatics*, 2014. ISSN 15320464. doi: 10.1016/j.jbi.2013.11.003.

Alan R. Aronson and François Michel Lang. An overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 2010. ISSN 10675027. doi: 10.1136/jamia.2009.002733.

Olivier Bodenreider. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 2004. ISSN 03051048. doi: 10.1093/nar/gkh061.

Wendy W. Chapman, Prakash M. Nadkarni, Lynette Hirschman, Leonard W. D'Avolio, Guergana K. Savova, and Ozlem Uzuner. Overcoming barriers to NLP for clinical text: The role of shared tasks and the need for additional creative solutions, 2011. ISSN 10675027.

Yan Chen, Huanying (Helen) Gu, Yehoshua Perl, and James Geller. Structural group-based auditing of missing hierarchical relationships in UMLS. *Journal of Biomedical Informatics*, 2009. ISSN 15320464. doi: 10.1016/j.jbi.2008.08.006.

J. J. Cimino, H. Min, and Y. Perl. Consistency across the hierarchies of the UMLS Semantic Network and Metathesaurus. *Journal of Biomedical Informatics*, 2003. ISSN 15320464. doi: 10.1016/j.jbi.2003.11.001.

Dina Demner-Fushman, Wendy W. Chapman, and Clement J. McDonald. What can natural language processing do for clinical decision support?, 2009. ISSN 15320464.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. 2010. doi: 10.1109/cvpr.2009.5206848.

Noémie Elhadad, Sameer Pradhan, Sharon Gorman, Suresh Manandhar, Wendy Chapman, and Guergana Savova. SemEval-2015 Task 14: Analysis of Clinical Text. 2015. doi: 10.18653/v1/s15-2051.

Paolo Ferragina and Ugo Scaiella. TAGME: On-the-fly annotation of short text fragments (by Wikipedia entities). In *International Conference on Information and Knowledge Management, Proceedings*, 2010. ISBN 9781450300995. doi: 10.1145/1871437.1871689.

Carol Friedman, Lyudmila Shagina, Yves Lussier, and George Hripcsak. Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*, 2004. ISSN 10675027. doi: 10.1197/jamia.M1552.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare, 2019. ISSN 1546170X.

Michael Halper, C. Paul Morrey, Yan Chen, Gai Elhanan, George Hripcsak, and Yehoshua Perl. Auditing hierarchical cycles to locate other inconsistencies in the UMLS. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2011. ISSN 1942597X.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2011. ISBN 1937284115.

Peter B. Jensen, Lars J. Jensen, and Soøren Brunak. Mining electronic health records: Towards better research applications and clinical care, 2012. ISSN 14710056.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 2016. ISSN 20524463. doi: 10.1038/sdata.2016.35.

Nikolaos Kolitsas, Octavian Eugen Ganea, and Thomas Hofmann. End-to-end neural entity linking. In *CoNLL 2018 - 22nd Conference on Computational Natural Language Learning, Proceedings*, 2018. ISBN 9781948087728. doi: 10.18653/v1/k18-1050.

Yen Fu Luo, Weiyi Sun, and Anna Rumshisky. MCN: A comprehensive corpus for medical concept normalization, 2019. ISSN 15320464.

Prakash M. Nadkarni, Lucila Ohno-Machado, and Wendy W. Chapman. Natural language processing: An introduction, 2011. ISSN 10675027.

Stuart J. Nelson, Kelly Zeng, John Kilbourne, Tammy Powell, and Robin Moore. Normalized names for clinical drugs: RxNorm at 6 years. *Journal of the American Medical Informatics Association*, 2011. ISSN 10675027. doi: 10.1136/amiajnl-2011-000116.

NIH-NLM. SNOMED Clinical Terms® (SNOMED CT®), 2015.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. *ACM CHIL*, 2020. URL http://arxiv.org/abs/1909.12475.

Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. SemEval-2014 Task 7: Analysis of Clinical Text. 2015. doi: 10.3115/v1/s14-2007.

Jonathan Raiman and Olivier Raiman. DeepType: Multilingual entity linking by neural type system evolution. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 2018. ISBN 9781577358008.

Ruth Reátegui and Sylvie Ratté. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. *BMC Medical Informatics and Decision Making*, 2018. ISSN 14726947. doi: 10.1186/s12911-018-0654-2.

Alejandro Riveros, Maria De Arteaga, Fabio González, Sergio Jimenez, and Henning Müller. MindLab-UNAL: Comparing Metamap and T-mapper for Medical Concept Extraction in SemEval 2014 Task 7. 2015. doi: 10.3115/v1/s14-2073.

Maya Rotmensch, Yoni Halpern, Abdulhakim Tlimat, Steven Horng, and David Sontag. Learning a Health Knowledge Graph from Electronic Medical Records. *Scientific Reports*, 2017. ISSN 20452322. doi: 10.1038/s41598-017-05778-z.

Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 2010. ISSN 10675027. doi: 10.1136/jamia.2009.001560.

Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 2019. ISSN 1527974X. doi: 10.1093/jamia/ocz096.

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, Gareth J.F. Jones, Johannes Leveling, Liadh Kelly, Lorraine Goeuriot, David Martinez, and Guido Zuccon. Overview of the ShARe/CLEF eHealth evaluation lab 2013. *Lecture Notes in Computer Science*, 8138 LNCS:212–231, 2013. ISSN 03029743. doi: 10.1007/978-3-642-40802-1_24.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 2011. ISSN 10675027. doi: 10.1136/amiajnl-2011-000203.

Yonghui Wu, Joshua C. Denny, S. Trent Rosenbloom, Randolph A. Miller, Dario A. Giuse, and Hua Xu. A comparative study of current Clinical Natural Language Processing systems on handling abbreviations in discharge summaries. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2012. ISSN 1942597X.

Appendix

Appendix A: Description of Annotation Tool Suggestion System

Below, we describe the lightweight suggestion system that underlies our annotation tool. Due to its simplicity, it can run seamlessly. First, we constructed two data structures from UMLS tables, using only the terms from SNOMED and RxNorm.

- A lookup table. This is a dictionary mapping text to all of the CUIs that have that text as a Preferred Name or Synonym.

- An inverted index. The inverted index was built as per classical information retrieval. The 'document' for a CUI was a concatenation of stemmed versions of all the words in that CUI's Preferred Name or Synonym.

The suggestions then consist of:

- Any direct matches of the text span, found in the lookup. They are sorted by the number of synonyms they have, which we use as a proxy for population prevalence.

- The closest set of matches from the inverted index. Before searching, we remove stop words from the highlighted text and stem the remainder of the words. Among equally relevant matches, we again sort by number of synonyms as a proxy for population prevalence.

Appendix B: Performance of Systems by Semantic Type

Below we show the performance of the top 10 systems of the n2c2 Shared Task on the test set, subdivided by the 20 most common Semantic Types. We note that a CUI can have multiple Semantic Types, in which case it is included in several calculations.

Test Subset	# of Examples	Max Accuracy	Avg Accuracy	Pooled Accuracy
Disease or Syndrome	713	88.8%	85.7%	96.2%
Finding	674	78.8%	75.1%	86.4%
Pharmacologic Substance	667	93.9%	91.9%	97.2%
Laboratory Procedure	665	88.7%	85.8%	94.4%
Organic Chemical	624	95.5%	94.4%	98.1%
Therapeutic or Preventative Procedure	565	77.5%	72.7%	89.4%
Sign or Symptom	456	94.3%	92.3%	96.9%
Diagnostic Procedure	411	90.3%	84.6%	95.4%
Qualitative Concept	311	94.9%	89.0%	96.8%
Health Care Activity	299	84.6%	81.6%	89.3%
Spatial Concept	241	92.5%	89.0%	95.4%
Pathologic Function	228	89.0%	85.8%	94.7%
Body Part, Organ, or Organ Component	189	82.0%	65.2%	89.9%
Quantitative Concept	165	95.8%	92.8%	97.0%
Temporal Concept	122	85.2%	83.6%	91.8%
Medical Device	108	78.7%	70.0%	88.0%
Antibiotic	102	97.1%	95.6%	98.0%
Neoplastic Process	100	91.0%	84.8%	97.0%
Functional Concept	93	87.1%	77.0%	88.2%
Amino Acid, Peptide, or Protein	71	91.5%	86.9%	91.5%

Appendix C: Full Clinical Annotation Schema Guidelines

Overall Guidelines

Tag all words that seem like medical terms, i.e. they require medical knowledge to understand or are frequently used in medicine.

Concepts and Actions (Nouns and Verbs)

- Tag all labs, e.g. "WBC count", "hematocrit", or "urinalysis".

- Tag all procedures, e.g. "EKG", "chest x-ray", "fingerstick", or "CT".

- Tag all treatments or interventions, e.g. "transfusion", "aspirin", "CABG", or "surgery".

  - Tag interventions even if they're common non-medical concepts, e.g. "fluids" or "tube".

  - Tag normal English verbs that are used to describe an intervention, e.g. for "blood cultures were drawn" tag "drawn" as "sample obtained".

  - For drugs, tag the drug name and dosage route, but do not tag the dosage. E.g. do tag "PO" but do not tag "15mg".

- Tag all symptoms, e.g. "headache" or "nausea".

  - Tag symptoms even if they are not medical terms, i.e. for "chest pain" tag "chest", "pain", and "chest pain".

- Tag all findings, e.g. "respiratory distress", "edema", "blood pressure", or "heart rate".

- Tag all diagnoses, e.g. "COPD", "MI", or "osteoporosis".

- Tag all medical history elements, e.g. "trauma" or "smoking history".

- Tag all healthcare terms, e.g. "PCP", "ED", "ambulance", "EMR", "OSH", "cardiologist", "rheum" (or any other medical specialty), "follow-up", "review of systems", "past medical history", "patient", "assessment", or "consult".

- Tag all illnesses or medical findings, e.g. "injury".

- Do not tag negation words, e.g. for "denies chest pain" do not tag "denies".

Modifiers (Adjectives and Adverbs)

- Tag all words that modify a medical concept.

- Tag all medical modifiers, e.g. "acute", "pleuritic", or "friable".

- Tag all anatomical terms, any words that refer to body positioning, e.g. "orbital", "right", or for "chest x-ray" tag "chest".

- Tag all descriptions of findings or a patient's state, e.g. "febrile" or "stable".

    - Do not tag adjectives that describe something medical but do not have a separate medical meaning from their English meaning, e.g. for "vital signs normal" do not tag "normal".

- Tag all descriptions of diseases or diagnoses that are considered medical terms or have separate medical meanings, e.g. "acute", "chronic", "mild", "moderate", "severe", "dynamic".

    - For example, tag "sensitive" if it is in the context of antimicrobial sensitivity, e.g. for "blood cultures grew E. Coli sensitive to x, y, z" tag "sensitive" because it describes a medical concept of antimicrobial sensitivity.
    - Do not tag a term if you search for it and there's no tag available, e.g. "profuse".
    - Do not tag the words "likely", "unlikely", or "possible".

References to Parts of the Note

- Tag references to parts of notes even though they are English words, but do not tag individual words that comprise them.

    - For example, "HPI" (History of Present Illness), "PMH" (Past Medical History), "A/P" (Assessment and Plan)

Span Splitting

- Tag sub-concepts if they have their own medical meaning, e.g. for "extraocular movements intact" tag "extraocular", "extraocular movements", and "extraocular movements intact". However, in the case of "respiratory failure" the word "failure" does not have any separate medical, anatomical, or physiological relevance so it would not be tagged.

- Do not tag sub-concepts if they are contained within other concepts that express the same meaning more fully written out, e.g. for "Crohn's disease" tag "Crohn's disease" and "disease" but not "Crohn's".

- Do not tag sub-spans of very common administrative terms used to describe portions of a note, e.g. for "history of present illness" do not tag "history", "present", or "illness".

Multiple Concept Matches

- Tag multiple CUIs when there are multiple that appear to be a good match in context, e.g. for "fever" potentially tag "fever symptom (finding)" and "fever (sign/symptom)".

- Even if a term can have multiple taggable meanings, only tag the meaning in context.

- If two terms match equally well, only tag the verbatim text, e.g. for "acute kidney injury" only use tag "acute kidney injury" but not "kidney failure, acute".

No Concept Matches

- If there is no exact match in SNOMED, tag as "CUI-less" (and optionally any close approximations)

  - If both subconcepts in a compound concept are tagged, there is no need to tag the compound concept as "CUI-less."

  - Do not use "CUI-less" to tag terms that are only semantically off (the concept "patients" for "patient").

- If there is an approximate match, feel free to tag as both "CUI-less" and those close approximations, e.g. "presentation" for "presents with".

Special Considerations

- If a term repeats within the same note, tag the term every time it appears.

- Ignore typos and tag the term as its intended meaning.

- Tag common medical abbreviations, e.g. "h/o", "PMH", "ggt", or "p/w".

  - Do not split abbreviations into different spans, even if they represent different medical concepts, e.g. for "EOMI" do not tag "EO" and "EOM" separately.

- For simplicity, tag lab tests with priority order level (aka finding) > measurement (procedure) > substance, e.g. for "albumin level" tag with "albumin level measurement" instead of "albumin". While multiple may be equally valid, this can be fixed in post-processing.

  - If a phrase describes the results of a lab test, only tag the substance tested itself and not the modifier, e.g. for "low albumin" tag "albumin" and do not tag "low".

  - Similarly, do not tag the numeric value if given, e.g. for "INR 3.1" do not tag "3.1".

  - Do tag lab test results with modifiers if the resulting state has a name, e.g. for "elevated WBC count" tag the phrase as "leukocytosis".

- Do not tag units of measurement.