Published in final edited form as:
J Biomed Inform. 2015 October ; 57: 28–37. doi:10.1016/j.jbi.2015.07.010.

Challenges in Clinical Natural Language Processing for Automated Disorder Normalization

Robert Leaman[1], **Ritu Khare**[1], and **Zhiyong Lu**[1,*]

Robert Leaman: robert.leaman@nih.gov; Ritu Khare: ritu.khare@nih.gov; Zhiyong Lu: zhiyong.lu@nih.gov

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, 20894

Abstract

Background—Identifying key variables such as disorders within the clinical narratives in electronic health records has wide-ranging applications within clinical practice and biomedical research. Previous research has demonstrated reduced performance of disorder named entity recognition (NER) and normalization (or grounding) in clinical narratives than in biomedical publications. In this work, we aim to identify the cause for this performance difference and introduce general solutions.

Methods—We use closure properties to compare the richness of the vocabulary in clinical narrative text to biomedical publications. We approach both disorder NER and normalization using machine learning methodologies. Our NER methodology is based on linear-chain conditional random fields with a rich feature approach, and we introduce several improvements to enhance the lexical knowledge of the NER system. Our normalization method – never previously applied to clinical data - uses pairwise learning to rank to automatically learn term variation directly from the training data.

Results—We find that while the size of the overall vocabulary is similar between clinical narrative and biomedical publications, clinical narrative uses a richer terminology to describe disorders than publications. We apply our system, DNorm-C, to locate disorder mentions and in the clinical narratives from the recent ShARe/CLEF eHealth Task. For NER (strict span-only), our system achieves precision = 0.797, recall = 0.713, f-score = 0.753. For the normalization task (strict span + concept) it achieves precision = 0.712, recall = 0.637, f-score = 0.672. The improvements described in this article increase the NER f-score by 0.039 and the normalization f-score by 0.036. We also describe a high recall version of the NER, which increases the normalization recall to as high as 0.744, albeit with reduced precision.

Discussion—We perform an error analysis, demonstrating that NER errors outnumber normalization errors by more than 4-to-1. Abbreviations and acronyms are found to be frequent

[*]Corresponding author: Zhiyong Lu, zhiyong.lu@nih.gov, MSC3825 NCBI/NLM/NIH, Bldg 38A, Rm 1003A, 8600 Rockville Pike, Bethesda, MD, 20894, Tel: 301-594-7089.

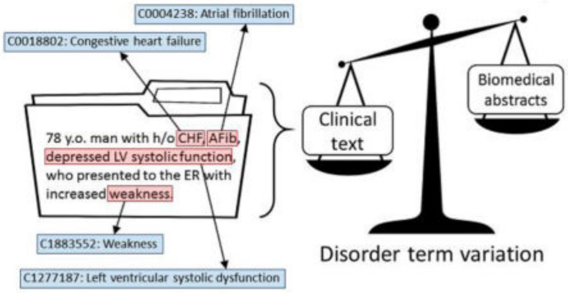
Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

causes of error, in addition to the mentions the annotators were not able to identify within the scope of the controlled vocabulary.

Conclusion—Disorder mentions in text from clinical narratives use a rich vocabulary that results in high term variation, which we believe to be one of the primary causes of reduced performance in clinical narrative. We show that pairwise learning to rank offers high performance in this context, and introduce several lexical enhancements – generalizable to other clinical NER tasks – that improve the ability of the NER system to handle this variation. DNorm-C is a high performing, open source system for disorders in clinical text, and a promising step towards NER and normalization methods that are trainable to a wide variety of domains and entities.

DNorm-C is open source software, and is available with a trained model at the DNorm demonstration website: http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/tmTools/#DNorm.

Graphical Abstract


C0004238: Atrial fibrillation
C0018802: Congestive heart failure
78 y.o. man with h/o CHF, AFib, depressed LV systolic function, who presented to the ER with increased weakness.
C1883552: Weakness
C1277187: Left ventricular systolic dysfunction
Clinical text
Biomedical abstracts
Disorder term variation

Keywords

natural language processing; electronic health records; information extraction

Introduction

The application of clinical natural language processing to the clinical narratives in electronic health records has the potential to make a significant impact in many aspects of healthcare and biomedical research. Recent work has demonstrated this potential at the point of care (such as with clinical decision support [1, 2]), for patients understanding their own records [3], for public health (e.g. biosurveillance [4]), and in biomedical research (e.g. cohort identification [5, 6], identifying novel potential clinical associations [7, 8] or pharmacovigilance [9, 10]).

A common task in clinical natural language processing is the identification of key clinical variables mentioned in the text, such as disorders or treatments. This subtask includes both locating mentions of key entities (named entity recognition) and identifying mentions found (normalization) with respect to a controlled vocabulary such as SNOMED-CT [11]. An example sentence from a clinical narrative, together with annotations for disorder mentions and their respective concepts can be seen in Figure 1. The results of these subtasks are then used by downstream components to provide the higher level processing required by the end task, making the result highly dependent on the quality of the normalization results obtained.

While clinical natural language processing has been an area of increasing attention, progress still trails the general and biomedical domains [12]. While this is partially due to the relative scarcity of clinical narrative corpora due to privacy concerns, it is also partially due to the additional difficulties encountered in clinical narrative text. Biomedical text is written to communicate research results to a wide audience and is edited for clarity. Clinical narrative text, on the other hand, is written by health care professionals to communicate the status and history of a single patient to other health care professionals or themselves. These notes record the reason the patient was seen (i.e. the chief complaint), document treatments given, the findings of any tests administered, and other information needed to make informed decisions regarding care for a single patient. Typical notes include progress notes documenting treatment provided, consult reports communicating the results of diagnostic tests, and discharge summaries recording the entire course of a hospital stay. As these notes each have different purposes, they are highly heterogeneous in their content and level of detail. Moreover their format is typically unstructured, resulting in a wide variety of ad-hoc structural components.

We illustrate many of the problems clinical narrative text presents for natural language processing with examples in Table 1. In general, systems which process clinical narratives cannot expect the clear structure and communication common in published text. Instead, clinical narrative is prepared under considerable time pressure, using a combination of ad-hoc formatting, eliding words which could be inferred from context, and with liberal use of parenthetical expressions, jargon and acronyms to increase the information density.

In this article, we approach the task of normalization of key entities in clinical narrative using disorders as a case study. While disease normalization in biomedical publications is a difficult task due to the wide variety of naming patterns for diseases, ambiguity, and term variation, the idiosyncrasies of clinical text cause additional difficulties. Disorder normalization in clinical text has been previously attempted using both lexical and rule-based techniques [13–17]. In this line of research, the task has often been to identify diagnoses, as in ICD-9 coding [18], or only a specific subset of problems [19]. In this study, we instead describe a system to identify and normalize all disorders mentioned in a clinical narrative. We employ DNorm, a machine learning method for learning the similarity between mentions and concept names from a specified controlled vocabulary directly from training data. Previous work with DNorm demonstrated state-of-the-art performance in biomedical publications [20]. Our participation in the recent ShARe/CLEF eHealth shared task resulted in the highest normalization performance of any participant [3], but also demonstrated performance significantly reduced compared to biomedical publications. In this manuscript we demonstrate that that the vocabulary used to describe disorders in clinical text is richer than in publications – and hence more difficult to model – while the variability in overall language use is similar (see Methods Section). We use this insight to make improvements to the named entity recognition component to handle the richer disorder vocabulary.

Related work

Natural language processing of clinical text has a long history, going back to the Linguistic String Project – Medical Language Processor (LSP-MLP) project in 1986 [21]. The field was thoroughly surveyed by Meystre et. al. [22] in 2008.

A recurring theme is the scarcity of annotated corpora, or datasets which can be used to develop and evaluate natural language processing systems [12]. One consequence of the relative lack of annotated data is a longstanding emphasis on knowledge intensive approaches. For example, two of the most widely used tools (MetaMap [23, 24] and MedLEE [25]) both emphasize a hybrid of natural language processing and lexical approaches using the UMLS Metathesuarus, rather than machine learning methods trained directly on clinical text. MetaMap, created by the National Library of Medicine, has been successfully applied to clinical narratives for a wide variety of purposes, including biosurveillance [4], cohort identification [5], and to find potentially novel clinical associations [7]. MedLEE has also been applied to clinical text for a wide variety of purposes, including pharmacovigilance [9], SNOMED-CT coding [25], and determining comorbidities [26].

The field has moved forward in large part due to the efforts of several groups to provide annotated data through the context of a shared task. The 2007 Medical NLP challenge involved assigning ICD-9 codes to radiology reports [18]. The 2010 i2b2 obesity challenge consisted of predicting the status of obesity and 15 related comorbidities, requiring participants to determine whether the disorders were mentioned [19]. This dataset was recently used by Tang et. al. [27] to compare several techniques for named entity recognition of disorders in hospital discharge summaries. The 2012 i2b2 challenge required the identification of clinically significant events (including clinical problems) and their temporal relationships [28].

One recent shared task specifically assessed the state of the art on automatic disorder recognition in clinical narrative text. This task, called ShARe/CLEF eHealth Task 1 [3], required participants to locate the span of all disorders mentioned in the clinical narrative (Task 1a) – the task of named entity recognition (NER) – and normalize (or ground) the mention to a concept within the Disorder semantic group of the controlled vocabulary SNOMED-CT (Task 1b).

A notable trend that can be seen in the shared tasks is a shift in from lexical and natural language processing based techniques for (NER) to techniques based primarily on machine learning. In the 2010 i2b2 challenge, for example, the highest-performing systems used a lexical approach, while in the 2012 challenge, the highest-performing systems were hybrids employing machine learning for NER and rule-based approaches for normalization (or grounding). Most methods for automatic normalization of diseases and related concepts from free text continue to rely on dictionary methods and/or heuristic rules [13–17].

We recently introduced a novel machine learning framework for normalization called DNorm [20], which showed state-of-the-art performance on the NCBI Disease Corpus, which is comprised of the abstracts of biomedical publications [29, 30]. This framework uses pairwise learning-to-rank to directly learn a similarity function between mentions found in narrative text and terms from a controlled vocabulary. Since the method is used with machine learning methods for NER – such as conditional random fields with a rich feature set – it provides an adaptable system that is quickly trainable to specific needs. While our participation in the ShARe/CLEF eHealth Tasks 1a and 1b [31] resulted in the highest normalization performance of any participant, it also demonstrated that the performance of NER and normalization in the clinical domain is significantly lower than in biomedical publications. In this article we identify reduced NER performance due to increased term variation as the primary cause for this performance difference, and introduce several improvements that are generalizable to other clinical NER problems to enhance the lexical knowledge of the NER system.

Methods

We created a pipeline system employing a statistical named entity recognizer, BANNER [32], and adapted a recently developed normalization method, DNorm [33], to perform the normalization. We perform an analysis demonstrating that term variation is a primary concern in clinical narrative, and use this to motivate additional lexical features in BANNER. Since DNorm learns term variants directly from training data that must be annotated for both mention span and concepts, we employed the ShARe/CLEF eHealth Task 1 corpus [3], which is annotated with disorder concepts from SNOMED-CT [11], a large structured vocabulary of clinical terms. We leveraged terms from the full UMLS Metathesaurus [34] to prepare the lexicon. This section describes each step in detail.

ShARe/CLEF eHealth Task 1 Corpus

The ShARe/CLEF eHealth Task 1 Corpus (the “eHealth corpus”) consists of 299 clinical notes (199 for training, and 100 for test) that were independently annotated by two professional trained annotators, followed by an adjudication step, resulting in a high inter-annotator agreement [3]. The organizers did not report harmonizing the annotations across the corpus. Annotations in this corpus represent disorders and are marked with the associated textual span and concept identifier from SNOMED-CT [11]. The annotators defined ‘disorder’ as “any span of text which can be mapped to a concept in SNOMED-CT and which belongs to the Disorder semantic group” [3], where the Disorder semantic group was defined to be all semantic types in the UMLS Disorders Semantic Group except Findings, a total of 11 types (Congenital Abnormality, Acquired Abnormality, Injury or Poisoning, Pathologic Function, Disease or Syndrome, Mental or Behavioral Dysfunction, Cell or Molecular Dysfunction, Experimental Model of Disease, Signs and Symptoms, Anatomical Abnormality, Neoplastic Process). The concept identifier used is the unique identifier used in the version of SNOMED-CT present in the UMLS Metathesaurus: the CUI, or Concept Unique Identifier. The training set contains 5784 disorder annotations and the test set contains 5351 disorder annotations.

The corpus contains four types of clinical notes: discharge summaries, electrocardiogram (ECG) reports, echocardiogram (echo) reports, and radiology reports, as shown in Table 2.

While the number of each type is relatively similar in the Training set, the Test set is dominated by discharge summaries, which are significantly longer than the other types.

Example sentences from the corpus are shown in Table 3, illustrating some distinctive characteristics of this corpus. The first example illustrates a single span which exactly matches the name of a SNOMED-CT concept, after case normalization. The term variation examples depict the case where the mention text refers to a SNOMED-CT concept via a name different than the one present in SNOMED-CT. Abbreviations, the next example, are an important sub-type of term variation, and – like term variations in general – may be unique to the physician responsible for the clinical narrative. The morphology example shows that term variations may consist of word variations rather than word substitutions. The next example demonstrates a “CUI-less” mention: a span considered to represent a disorder by the annotators, but which could not be resolved to a suitable concept within the SNOMED-CT Disorder group. Mentions annotated as “CUI-less” comprise 28.2% of all annotations in this corpus – a frequency somewhat less than other studies [35]. While these mentions cannot be normalized to a concept, their mention text contains useful, meaningful information that would be ignored if “CUI-less” mentions were disregarded. The disjoint mention example shows that disorder annotations may consist of spans which are not contiguous. Among the 11.1% of the annotations which are disjoint, 93.2% have 2 spans and the remainder have 3. The final example shows that disorders referenced indirectly are not annotated. A glucose reading of far over 180 strongly suggests diabetes, but this requires inference and is therefore not annotated.

Preliminary Corpus Analysis

We performed a comparative analysis of the NCBI Disease Corpus against the eHealth corpus to guide our adaptation of DNorm from the former to the latter. Because DNorm uses machine learning methods, the performance depends on how well the patterns learned from the training set generalize to the full domain. In a concept identification task such as disorder normalization, a primary concern is the size of the vocabulary of the domain: as the richness of the vocabulary increases, small amounts of text – such as the training data – become less representative of the full domain, causing performance to decrease. We quantified the notion of vocabulary richness using closure properties. A textual domain is said to exhibit closure with respect to some linguistic phenomenon if the total number of distinguishable elements of that aspect – as observed in a representative text – trends toward a finite number [36]. Closure properties can be analyzed visually by plotting the number of unique elements against the total number of elements over samples of increasing sizes. As this analysis guides the work adapting DNorm to clinical text, we describe both the analysis method and its results here.

Our first comparison measures the richness of the overall vocabulary by considering the closure of all tokens. We randomly shuffled the sentences in the training sets of both the eHealth and the NCBI Disease corpora. We then iterated through all sentences, sampling the number of unique tokens seen at regular intervals of the total number of tokens seen. We repeated this procedure 30 times, calculating the mean for each sample. The results for both corpora are seen in Figure 2. We see that the closure curves are very similar between the two corpora, implying that – when all tokens are considered – the richness of the vocabulary across the two domains is similar. Thus the reduced performance in the clinical domain cannot be explained by difference in the richness of the overall vocabulary.

We next compared the closure of mention texts and concepts. We randomly shuffled the mentions in the training sets of both the eHealth and the NCBI Disease corpora and iterated through all mentions, sampling the number of unique mention texts and unique concept identifiers seen at regular intervals of the total number of mentions considered. We repeated this procedure 30 times, calculating the mean for each sample. The results are seen in Figure 3, which demonstrates a much wider variety of both mention texts and concepts in clinical narrative compared to biomedical publications. Thus, while clinical narrative and biomedical publications have similar variability when considering all tokens, clinical narrative uses a richer vocabulary to describe disorders than publications. This higher variability makes inference within the domain more difficult and explains the reduced performance in clinical narrative text. Compensating for the increased variability of text in the clinical narrative thus becomes the primary goal of adapting DNorm to clinical narrative.

Lexicon Description

The lexicon is a pre-specified set of concepts, each with a unique identifier and one or more names. SNOMED-CT forms the core of our lexicon, since all annotations in the eHealth corpus are to SNOMED-CT concepts. Our initial lexicon was derived from the 2012AB release of the UMLS Metathesaurus [34]. The concepts included in the lexicon consisted of all those contained in the SNOMED-CT source vocabulary and belonging to the same 11 semantic types as used in the corpus, that is, the Disorder semantic group except for Findings. We increased the number of synonymous terms for each concept by including all English terms for each included concept except those marked suppressed, regardless of the source vocabulary for the term.

The initial lexicon derived from the UMLS Metathesaurus did not handle “CUI-less” mentions, even though these constitute over 30% of the corpus. Manual analysis revealed that many of the “CUI-less” concepts corresponded to concepts within the semantic type Findings. We determined, however, that simply including the Findings semantic type within the lexicon would significantly reduce precision. Instead, we identified the “CUI-less” mentions occurring five or more times in the Training set, and appended those mentions to the lexicon as terms for the concept “CUI-less.”

Further analysis of the Training set demonstrated that many noun forms of a term were freely substituted with the corresponding adjective form. While our system employed stemming, as described in the next section, stemming does not handle derivational morphology, and we found anatomical terms to be a particular concern. For example, the adjective form of “femur” is “femoral,” which is stemmed to “femor,” and different bases are also occasionally used, such as “optic” used as the adjective form of “eye.” We handled these by extracting a list of anatomic adjective/noun pairs from the UMLS (about 60), and adding synonyms to the lexicon substituting the adjective form for every lexicon name containing the noun form.

The Training set also contained several abbreviations that are not found in the Metathesaurus. To address these, we used the medical abbreviations from Taber’s Medical Dictionary [37]. We filtered the list to only include those entries where the expanded form exactly matched one of the terms for an included concept, which added 102 terms to the lexicon.

Finally, we noted that several abbreviation mentions in the Training set were ambiguous, e.g., the mention “AR” matches abbreviations listed in the UMLS for both “aortic regurgitation” (CUI C0003504) as well as “rheumatoid arthritis” (CUI C0003873), and “CAD” matches abbreviations for both “coronary heart disease” (CUI C0010068) as well as “coronary artery disease” (CUI C1956346). We filtered abbreviations from the lexicon where we could not match each letter in the abbreviation against the first letters of any of the other terms for the same concept. E.g. “AR” would match “aortic regurgitation” but not “rheumatoid arthritis.” Some abbreviations required additional disambiguation e.g., “MI” matches both “myocardial infarction” as well as “mitral incompetence.” We resolved these cases by preferring the sense that appears more frequently in the Training set.

System Description

DNorm includes separate modules for mention-level NER using BANNER [32], abbreviation resolution, and normalization using a pairwise learning-to-rank algorithm. In this study, we adapt DNorm to clinical notes by introducing additional NER features and a post-processing module to handle the increased term variation. An overview of the overall approach is provided in Figure 4, and described in greater detail in the subsequent sections.

Preprocessing

While clinical narratives are structured, the structure varies between different types of reports and different clinical groups, with the structure variably applied even within these subsets [38]. Because local context is more important than global context for inferring whether a specified span refers to a disorder mention, we segment each clinical narrative into individual sentences. Sentence segmentation in clinical narrative is complicated by two factors, however. First, it is not always clear what constitutes a sentence in the clinical narrative (see, e.g., the “non-sentential structure” examples in Table 1). Second, sentence punctuation is used inconsistently, leading to some sentences lacking periods, or lists which lack separating commas or newlines.

Our preprocessing module initially segments the text using the implementation of sentence segmentation in the standard distribution of Java[1]. We then refine the segmentation with rules we identified manually by iteratively applying our module to the Training subset. These rules locate several additional sentence breaks, such as multiple adjacent newlines, and also correct some consistent errors, such as inserting a sentence break after “Dr.”

[1] https://docs.oracle.com/javase/7/docs/api/java/text/BreakIterator.html

Disorder mention recognition

We identified disorder mentions using the BANNER named entity recognition system [32]. BANNER is a trainable NER system which uses conditional random fields and a rich feature approach. We used an extensive feature set similar to previous work in NER for diseases [39], including the following:

- Individual tokens and lemmas: We included one binary feature for each token and lemma observed in the training set.
- Part of speech: We included a binary feature for each part of speech.
- Character n-grams: We included binary features for each character n-gram of length 2 and length 3 observed in the training set.
- Word shapes: We transformed each token by replacing each upper case letter with 'A', each lower case letter with 'a', each digit with '0', and all other characters with '_'. We included one binary feature for each transformed token observed in the training data. We also included a variant that replaces multiple characters of the same type with a single character.

We set the CRF order – the number of previous labels to be used to predict the next label – to 1. We initially set the labeling model to IO, which labels each token as either part of a disorder mention ("I") or not ("O"). This configuration does not effectively represent disjoint mentions, however, since it only models whether or not a token was a part of a mention. We therefore created a second model that employs different labels for contiguous and disjoint mentions (the CD label model). Examples of both the IO labels and the CD labels are seen in Table 4. Spans tagged by the model as contiguous mentions were returned directly, but spans labeled as part of a disjoint mention were joined into a single disjoint mention. This significantly reduced the confusion between contiguous and disjoint mentions, and allowed either 0 or 1 disjoint mentions to be accurately represented within each sentence. We note that while neither the IO nor the CD labeling models are capable of differentiating between adjacent mentions, these are relatively rare (2.4% of mentions in the training data) and introducing an additional label to delineate the start of a mention (as in the IOB model) increases the complexity of the patterns to be learned.

To improve the ability of the system to handle term variation, we performed several additional improvements to the NER component. First, we added the output of MetaMap [23, 24] as an input feature to the system ("MetaMap feature"). Given a textual passage, MetaMap identifies the candidate UMLS concepts and their corresponding mention spans. For this study, the source vocabulary was limited to SNOMED-CT, and the semantic categories were restricted to the 11 disorder semantic types as previously specified. The provided clinical report is split into chunks, and each chunk is fed into the MetaMap API to obtain the candidate mentions. Mentions which overlap are resolved to the longest span, e.g., "breast cancer" is preferred to "cancer." The module also filters some generic mentions, e.g., "allergies," "condition," "disease," "finding," etc.

Next, we improved the dictionary feature used by BANNER ("improved dictionary features") by adding the semantic type "Findings" and the other semantic types from the

Disorder subgroup not already included. We also added a second dictionary feature to locate anatomical mentions using the UMLS semantic type “Body Part, Organ or Organ Component.” These are useful since many disorder mentions include the part of the anatomy affected.

Normalization with DNorm

DNorm is a technique for learning the best mapping of a given mention to a name within a controlled vocabulary such as SNOMED-CT [20]. It converts both the mentions and the names from the controlled vocabulary to TF-IDF vectors [40] and then uses a regression model learned directly from the training data to score each name in the controlled vocabulary against the mention, and returns the top ranked name. We provide an overview of the method and describe the differences from previous implementations.

Both mentions and names to vectors are converted to lower case, each token is stemmed, stop words are removed, and the remaining tokens are converted to TF-IDF vectors [40]. The TF of each element in the vector is calculated as the number of times the corresponding token appears in the mention or name. The IDF for each element in the vector is calculated from lexicon as:

$$IDF=\log\frac{count(number\ of\ names\ in\ lexicon)}{count(number\ of\ names\ in\ lexicon\ containing\ the\ token)+1}$$

All vectors are normalized to unit length. We calculate the score between mention vector m and name vector n using a weight matrix W, whose elements W_{ij} encode the correlation between token m_i appearing in a mention and token n_j appearing in the concept name:

$$score(m,n)=m^T Wn=\sum_{i,j} m_i W_{ij} n_j$$

Under this framework, normalizing a given mention consists of finding the highest-ranking name and then returning its associated concept.

The matrix W is initialized to the identity matrix I, making the scoring function initially equivalent to cosine similarity [40]. W is then iteratively optimized via stochastic gradient descent [41]. Specifically, we iterate through each mention m from the training data, with its associated correct name n^+, and also iterate through each incorrect name n^-. If the score for $\langle m, n^+\rangle$ is not greater than the score for $\langle m, n^-\rangle$ by at least a constant margin (r) then the weight matrix W is adjusted by slightly increasing the correlation between m and n^+ and slightly decreasing the correlation between m and n^-. That is, if $m^T Wn^+ - m^T Wn^- < r$, then $W \leftarrow W + \lambda(m(n^+)^T - m(n^-)^T)$. We empirically determined that $\lambda = 10^{-3}$ provided the best performance.

In our previous work using DNorm in biomedical publications, where we employed the MEDIC vocabulary [42], we found that a margin of 1 ($r = 1$) provided better performance than a margin of 0 ($r = 0$) [20]. With the SNOMED-CT vocabulary, as used in this work, we found instead that a non-zero margin caused performance to drop significantly. We traced this issue to the SNOMED-CT vocabulary, which contains significantly more unique tokens than the MEDIC vocabulary but whose terms are also highly compositional, causing much of the vocabulary to be reused frequently [11]. The result of this compositionality is that using a margin of 1 with training mentions such as "fracture" causes the model to learn spurious negative correlations between "fracture" and the other tokens it appears with in the lexicon, such "femur." This, in turn, causes mentions employing these terms, such as "femur fracture," to be normalized incorrectly. Reducing the margin to $r = 0$ resolves these spurious negative correlations.

Post-processing

We implemented some rule-based post-processing to correctly handle several consistent patterns. For example, "w/r/r," is an abbreviation for "wheezing" (CUI C0043144), "rales" (CUI C0034642), and "ronchi" (CUI C0035508). We also included rules to handle common disjoint mentions, such as the physical exam finding "tender abdomen," and to filter some anatomical terms (e.g. "lung") which are false positives when they constitute the complete mention.

Results

Empirical feedback during system development was provided by reserving approximately 20% of the eHealth Training set for assessing improvements. Once development was complete, both the NER and the normalization models were retrained on the full Training set and evaluation was performed on the Test set, which was previously unseen.

We report the results of our experiments using multiple evaluation measures, all at the mention level, which evaluate the ability of the system to identify the correct disorder span and also the correct concept identifier. These consist of strict and relaxed versions of span-only precision, recall and F-score to evaluate NER, and strict and relaxed versions of span +concept precision, recall and F-score for evaluating normalization. Precision (*p*), recall (*r*), and F-score (*f*) are defined as follows:

$$p=\frac{tp}{tp+fp}\ r=\frac{tp}{tp+fn}\ f=\frac{2pr}{p+r}$$

where *tp* is defined as the number of spans that the system returns correctly; for the strict measure, all spans must match on both sides, the overlapping measure only requires the spans to share at least one token. For the span+concept measures, *tp* requires the correct concept identifier, in addition to either a strict or relaxed span match. All measures are micro-averaged. "CUI-less" mentions are evaluated as if "CUI-less" were their concept identifier (the system must return "CUI-less" or the concept will be marked incorrect).

Note that our definition of strict and relaxed span-only precision, recall and F-score are equivalent to the measures used in the ShARe/CLEF eHealth shared task, making these measurements directly comparable to those reported by the NER task (Task 1a) [3]. Moreover, the primary evaluation measurement of the normalization task (Task 1b), which the organizers called “strict accuracy,” is equivalent to our definition of strict span + concept recall, again allowing comparison between our results and those reported by the task.

The results of each evaluation measure are reported in Table 5 for the base version and each experimental condition. The base version is comparable to the version submitted to the eHealth task, but has slightly higher performance due to an update in the version of the UMLS used for the dictionary features. As the table shows, we have significantly improved results for both NER and normalization compared to the base version of the system: NER from a strict f-measure of 0.714 to 0.753 and normalization from a strict f-measure of 0.636 to 0.672. These improvements are statistically significant ($p < 0.01$; bootstrap resampling [41]). Furthermore, our final version (last row in Table 5) now has slightly higher performance than the highest NER strict F-score reported in the eHealth task (0.750). In addition, while our system previously had the highest normalization performance (as measured by strict recall) of the participants in the eHealth task, we have improved this measure by an additional 0.046.

Our system is able to provide significantly higher recall, at the expense of precision, through n-best decoding in the conditional random field model [42]. While conditional random field models are typically used to return the single highest scoring sequence of tags, in n-best decoding the n highest scoring sequences of tags are provided, typically resulting in several variations on the tagging sequence, increasing the probability that one is correct. The results of using n=5 and also n=50 are reported in Table 6 where we achieve a maximum strict span-only recall of 0.889, and a maximum strict span + concept recall of 0.744.

Discussion

Our approach is based on classification of the text using machine learning and local context. Several characteristics of the corpus contributed to our results. First, the annotators were instructed to annotate all disorders mentioned, even if not a current concern or not experienced by the patient, and also only annotate disorders that are referenced textually, rather than disorders requiring some inference. In addition, the annotators were requested to annotate spans that were an exact match for the concept being annotated. In particular, anaphoric references are not annotated and negation is ignored except for the uncommon case of a disorder described as the negation of a normal state, such as “not ambulated.” Since these characteristics all favor a local inference approach it is notable, therefore, that the results for both NER and for normalization remain lower than is typically reported for diseases in published abstracts, where reported performance ranges from 0.78 to 0.84 f-measure [20, 29, 30].

Error analysis

We manually analyzed the errors made by DNorm on the Test set, and report the results in this section. In this analysis an error could be either a false positive or a false negative. Normalization errors – a correct span but incorrect concept identifier – only constituted 17.2% of the errors. The remaining errors, 82.8%, are due to the NER (one or more incorrect spans). NER errors are therefore 4.8 times more frequent than normalization errors, underscoring the difficulty of NER for this task compared to normalization. This contrasts with published abstracts, where the NER errors were only 1.3 times more frequent than normalization errors [20]. This is particularly notable in light of the fact that the normalization task is significantly more difficult in clinical narrative since the vocabulary is larger and the MEDIC lexicon used for publications combines many near-synonym disease concepts into a single concept [42].

Analysis of the normalization errors revealed that most were between related concepts. These may be less specific than the annotated concept – such as returning C0023890 (“Liver cirrhosis”) for the mention “hepatitis-c cirrhosis,” instead of “CUI-less” – or may be more specific, such as returning C0013384 (“Dyskinetic syndrome”) for the mention “abnormal movements,” instead of C0392702 (“Abnormal involuntary movements”). We also found some annotations that appeared to be context-sensitive. For example, the mention “collapse” is annotated as “atelectasis” (CUI C0004144) when it refers to a collapsed lung but to “CUI-less” when it refers to other parts of the anatomy.

Due to the high frequency of NER errors (incorrect spans), we distinguished these into three further categories, which we describe here and then analyze below. *Boundary errors*, where the span(s) returned overlapped the correct span(s) without matching exactly, constitutes 27.6% of all errors. We also found 39.3% of all errors to be caused by *false negatives*: there was an annotated mention that was not matched (or overlapped) by any mention found. Finally, 16.0% of the total errors were due to *false positives*: the NER model returned a mention which did not overlap any annotated span.

Boundary errors frequently involved including too much detail (“apneic episodes” rather than just “apneic”) or too little (“chest pain” instead of “chest pain with exertion”). Errors involving disjoint mentions often fall into this category. Another frequent example is a list of medical conditions entered into the clinical note without delimiters, such as “Aspiration pneumonia Down’s syndrome Alzheimer’s dementia.” In 35.9% of the cases of boundary errors, the concept identifier assigned by DNorm was the same as the concept identifier annotated, suggesting that the evaluation is overly pessimistic in cases where usable normalization does not require the exact spans.

False negatives were frequently “CUI-less” mentions (57.6%). Also common were abbreviations (e.g. “DM2,” “OCR”), ad-hoc abbreviations (“extrem warm”), descriptive phrases (“food got stuck”), relatively generic problem adjectives (“narrowing,” “unsteady”) and nouns which do not clearly indicate a disorder out of context (“temperature,” “speech”).

False positives were frequently caused by confusion with entities of other types. A common example is treatments (“inflammatory medications,” “mitral valve repair”), which not only share vocabulary with disorders but can also be disorders themselves if abused (“opiates”). Other examples include mentions of disease-causing organisms (“Klebsiella pneumoniae”), bodily functions (“asleep”), and anatomical mentions (“umbilicus”).

Based on this initial analysis, summarized in Figure 5, we performed a second analysis to determine the overall frequency of several specific errors. These errors involved “CUI-less” mentions, disjoint mentions, abbreviations/acronyms, and misspellings.

First, we determined that either the annotated concept or the concept returned by DNorm was “CUI-less” for 50.2% of the errors. Contrasting this frequency with the frequency of “CUI-less” mentions in the corpus (28.2%) suggests that “CUI-less” mentions are more difficult to recognize, and this is a major source of error. Manual analysis of these mentions indicates that these are primarily caused by two phenomena: findings and hyponyms. Many mentions considered disorders by the annotators, such as “low blood pressure,” are listed in SNOMED-CT under the semantic type Finding (in this case, CUI C0020649, Hypotension). Since Findings fall outside the scope of the defined vocabulary, these were listed as “CUI-less.” Other “CUI-less” annotations, such as “hepatitis-c cirrhosis” are specializations of existing SNOMED-CT concepts, and therefore are annotated as “CUI-less.”

Second, we considered whether either the annotated span or the span found was disjoint. For example, the phrase “hemorrhagic contusion in the left frontal lobe” was annotated as “hemorrhagic contusion,” but the system returned the disjoint span “hemorrhagic contusion | left frontal lobe.” We found that disjoint spans were involved in 11.5% of the errors. While it seems reasonable that disjoint mentions would be more difficult to locate than contiguous mentions, comparing the similarity of their frequency in the set of errors to their frequency in the overall corpus (11.1%) suggests that our method of handling disjoint mentions was nearly sufficient to remove them as a performance bottleneck. In addition, we did see some evidence of the annotators using disjoint mentions to ensure that all tokens in the span annotated would match a token in the name for the annotated concept. For example, the phrase “mitral leaflets are mildly thickened” was annotated as “mitral leaflets | thickened.” This partially explains the relative complexity of the NER task compared to the normalization task.

Third, we considered whether either the annotated span or the span found involved abbreviations (e.g. “tachy”) or acronyms (e.g. “RA”). We found that these were involved in 21.7% of the errors, including both NER errors – from the difficulty in recognizing that the span represents a disease name – and normalization errors – from the difficulty in determining the correct concept. Acronyms and abbreviations therefore are a major source of error.

Fourth, we considered whether either the annotated span or the span found included a misspelling. Our analysis determined that misspellings typically caused false negatives in the normalization. For example, the misspelled disorder mention “pharyungitis” (for “pharyngitis”) retains sufficient hints – particularly the suffix “-itis” – that the NER model successfully infers that the token represents a disorder mention. The normalization model, on the other hand, only considers full tokens, and since the token is not present in the lexicon, normalization fails. While this is a limitation of our method, we found that misspelled words were only present in 1.9% of the errors. We therefore conclude that while misspellings are not a primary source of error in this corpus, addressing misspellings in future work would increase the generalization of the method.

Conclusion

In this work, we successfully adapted DNorm to locate and identify disorder mentions in clinical notes. Our comparative analysis of language use in clinical narrative and biomedical publications shows that there is no difference in the number of different words used in general between the two genera, though this result should be replicated with larger corpora. However our results show that clinical narrative text references a wider variety of disorders than biomedical publications and uses a larger number of phrases (mentions) to refer to them. This analysis provides an explanation for the performance differences between the two domains and motivated our inclusion of new lexical features in the NER and additional post-processing. DNorm uses a pipeline approach, chaining a trainable NER tool (BANNER) with a novel normalization methodology which learns term variations directly from the training data by applying a learning algorithm based on pairwise learning-to-rank. We demonstrated improved performance on both named entity recognition and normalization on the ShARe/CLEF eHealth corpus by improving both NER features and post-processing. Our analysis of the errors demonstrates that NER errors outnumber normalization errors by more than 4-to-1. This indicates that the NER task is more difficult than the normalization task, and suggests that future efforts be concentrated on NER. The error analysis shows that the primary sources of error are mentions that the annotators considered disorders but could not identify within the defined scope of the controlled vocabulary (the “CUI-less” mentions) and acronyms/abbreviations. We believe it may be possible to handle the “CUI-less” mentions by reclassifying out-of-scope UMLS concepts [44] and acronyms/abbreviations using graph-based disambiguation [45]. The error analysis also demonstrated that that misspellings are not a significant source of error. We believe that this method is widely applicable, and have now demonstrated its applicability in settings involving both clinical notes with mention-level normalization and published literature with document-level normalization [20].

Acknowledgments

The authors are grateful to the authors of the eHealth corpus, thank Dr. Chih-Hsuan Wei for his help preparing the demonstration website, and the anonymous reviewers for their insightful observations. This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

References

1. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform. 2009; 42:760–72. [PubMed: 19683066]
2. Pai VM, Rodgers M, Conroy R, Luo J, Zhou R, Seto B. Workshop on using natural language processing applications for enhancing clinical decision making: an executive summary. J Am Med Inform Assoc. 2014; 21:e2–5. [PubMed: 23921193]
3. Suominen, H.; Salanterä, S.; Velupillai, S.; Chapman, W.; Savova, G.; Elhadad, N., et al. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In: Forner, P.; Müller, H.; Paredes, R.; Rosso, P.; Stein, B., editors. Information Access Evaluation Multilinguality, Multimodality, and Visualization. Springer; Berlin Heidelberg: 2013. p. 212-31.
4. Chapman WW, Fiszman M, Dowling JN, Chapman BE, Rindflesch TC. Identifying respiratory findings in emergency department reports for biosurveillance using MetaMap. Studies in health technology and informatics. 2004; 107:487–91. [PubMed: 15360860]
5. Cui L, Sahoo SS, Lhatoo SD, Garg G, Rai P, Bozorgi A, et al. Complex epilepsy phenotype extraction from narrative clinical discharge summaries. J Biomed Inform. 2014

6. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc. 2014; 21:221–30. [PubMed: 24201027]

7. Hanauer DA, Saeed M, Zheng K, Mei Q, Shedden K, Aronson AR, et al. Applying MetaMap to Medline for identifying novel associations in a large clinical dataset: a feasibility analysis. J Am Med Inform Assoc. 2014

8. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nature reviews Genetics. 2012; 13:395–405.

9. Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. J Am Med Inform Assoc. 2009; 16:328–37. [PubMed: 19261932]

10. Iyer SV, Harpaz R, LePendu P, Bauer-Mehren A, Shah NH. Mining clinical text for signals of adverse drug-drug interactions. J Am Med Inform Assoc. 2014; 21:353–62. [PubMed: 24158091]

11. Stearns, MQ.; Price, C.; Spackman, KA.; Wang, AY. SNOMED Clinical Terms: Overview of the Development Process and Project Status. Proceedings of the AMIA Symposium; 2001. p. 662-6.

12. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. J Am Med Inform Assoc. 2011; 18:540–3. [PubMed: 21846785]

13. Gurulingappa, H.; Klinger, R.; Hofmann-Apitius, M.; Fluck, J. An Empirical Evaluation of Resources for the Identification of Diseases and Adverse Effects in Biomedical Literature. 2nd Workshop on Building and evaluating resources for biomedical text mining; Valetta, Malta. 2010. p. 15-21.

14. Névéol, A.; Kim, W.; Wilbur, WJ.; Lu, Z. Exploring two biomedical text genres for disease recognition. Proceedings of the ACL 2009 Workshop on Natural Language Processing in Biomedicine (BioNLP); 2009. p. 144-52.

15. Kang N, Singh B, Afzal Z, van Mulligen EM, Kors JA. Using rule-based natural language processing to improve disease normalization in biomedical text. Journal of the American Medical Informatics Association: JAMIA. 2012

16. Solt I, Tikk D, Gal V, Kardkovacs ZT. Semantic classification of diseases in discharge summaries using a context-aware rule-based classifier. J Am Med Inform Assoc. 2009; 16:580–4. [PubMed: 19390101]

17. Leaman, R.; Wojtulewicz, L.; Sullivan, R.; Skariah, A.; Yang, J.; Gonzalez, G. Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts to Health-Related Social Networks. Proceedings of the 2010 Workshop on Biomedical Natural Language Processing; Upsalla, Sweden. 2012. p. 117-25.

18. Pestian, JP.; Brew, C.; Matykiewicz, P.; Hovermale, DJ.; Johnson, N.; Cohen, KB., et al. A Shared Task Involving Multi-label Classification of Clinical Free Text. BioNLP 2007: Biological, translational and clinical language processing; Prague, Czech Republic. 2007. p. 97-104.

19. Uzuner O. Recognizing obesity and comorbidities in sparse data. J Am Med Inform Assoc. 2009; 16:561–70. [PubMed: 19390096]

20. Leaman R, Islamaj Dogan R, Lu Z. DNorm: Disease name normalization with pairwise learning-to-rank. Bioinformatics. 2013; 29:2909–17. [PubMed: 23969135]

21. Sager, N.; CF; EC. The analysis and processing of clinical narrative. In: Salamon, R.; Blum, B.; Jorgensen, M., editors. Medinfo 86. Amsterdam (Holland): Elsevier; 1986. p. 1101-5.

22. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearbook of medical informatics. 2008:128–44. [PubMed: 18660887]

23. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of the AMIA Symposium. 2001:17–21.

24. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010; 17:229–36. [PubMed: 20442139]

25. Lussier YA, Shagina L, Friedman C. Automating SNOMED coding using medical language understanding: a feasibility study. Proc AMIA Symp. 2001:418–22. [PubMed: 11825222]

26. Salmasian H, Freedberg DE, Friedman C. Deriving comorbidities from medical records using natural language processing. J Am Med Inform Assoc. 2013; 20:e239–42. [PubMed: 24177145]
27. Tang B, Cao H, Wu Y, Jiang M, Xu H. Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. BMC medical informatics and decision making. 2013; 13 (Suppl 1):S1. [PubMed: 23566040]
28. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc. 2013; 20:806–13. [PubMed: 23564629]
29. Doàan RI, Leaman R, Lu Z. NCBI disease corpus: A resource for disease name recognition and concept normalization. J Biomed Inform. 2014; 47:1–10. [PubMed: 24393765]
30. Doàan, RI.; Lu, Z. An improved corpus of disease mentions in PubMed citations. Proceedings of the ACL 2012 Workshop on BioNLP; 2012. p. 91-9.
31. Leaman, R.; Khare, R.; Lu, Z. NCBI at 2013 ShARe/CLEF eHealth Shared Task: Disorder Normalization in Clinical Notes with DNorm. Conference and Labs of the Evaluation Forum 2013 Working Notes; Valencia, Spain. 2013.
32. Leaman R, Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition. Pac Symp Biocomput. 2008:652–63. [PubMed: 18229723]
33. Leaman R, Doàan RI, Lu Z. DNorm: Disease name normalization with pairwise learning-to-rank. Bioinformatics. 2013; 29:2909–17. [PubMed: 23969135]
34. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004; 32:D267–70. [PubMed: 14681409]
35. Friedlin, J.; Overhage, M. An evaluation of the UMLS in representing corpus derived clinical concepts. AMIA Annual Symposium proceedings/AMIA Symposium AMIA Symposium; 2011. p. 435-44.
36. Temnikova, IP.; Cohen, KB. Recognizing sublanguages in scientific journal articles through closure properties. Workshop on Biomedical Natural Language Processing; Sofia, Bulgaria: Association for Computational Linguistics; 2013. p. 72-9.
37. Taber's Cyclopedic Medical Dictionary. 22. F.A. Davis Company; 2013.
38. Cohen, KB.; Demner-Fushman, D. Biomedical Natural Language Processing. John Benjamins Publishing Company; 2014.
39. Leaman R, Miller C, Gonzalez G. Enabling Recognition of Diseases in Biomedical Text with Machine Learning: Corpus and Benchmark. Proceedings of the 2009 Symposium on Languages in Biology and Medicine. 2009:82–9.
40. Manning, CD.; Raghavan, P.; Schütze, H. Introduction to Information Retrieval. Cambridge University Press; 2008.
41. Burges C, Shaked T, Renshaw E, Lazier A, Deeds M, Hamilton N, et al. Learning-to-rank using gradient descent. Proceedings of the International Conference on Machine Learning. 2005:89–96.
42. Davis AP, Wiegers TC, Rosenstein MC, Mattingly CJ. MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database. Database. 2012; 2012:bar065. [PubMed: 22434833]
43. Smith L, Tanabe LK, Ando RJ, Kuo CJ, Chung IF, Hsu CN, et al. Overview of BioCreative II gene mention recognition. Genome Biol. 2008; 9 (Suppl 2):S2. [PubMed: 18834493]
44. Fan JW, Friedman C. Semantic reclassification of the UMLS concepts. Bioinformatics. 2008; 24:1971–3. [PubMed: 18625612]
45. Chasin R, Rumshisky A, Uzuner O, Szolovits P. Word sense disambiguation in the clinical domain: a comparison of knowledge-rich and knowledge-poor unsupervised methods. J Am Med Inform Assoc. 2014

Highlights

- Disorder normalization in clinical text has wide-ranging applications.
- Clinical normalizers must handle ad-hoc formatting, jargon, and ambiguous acronyms.
- Disorder vocabulary is richer in clinical text than biomedical abstracts.
- Normalization with pairwise learning to rank handles rich vocabulary.
- Further normalization improvements require improved named entity recognition.

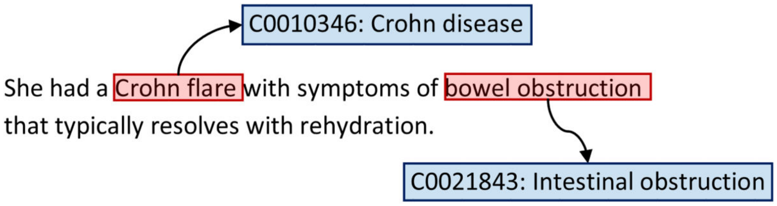

C0010346: Crohn disease
She had a Crohn flare with symptoms of bowel obstruction
that typically resolves with rehydration.
C0021843: Intestinal obstruction

Figure 1.
Example of normalization in clinical text, showing two disorder mentions, with their respective spans and SNOMED-CT concept identifiers.

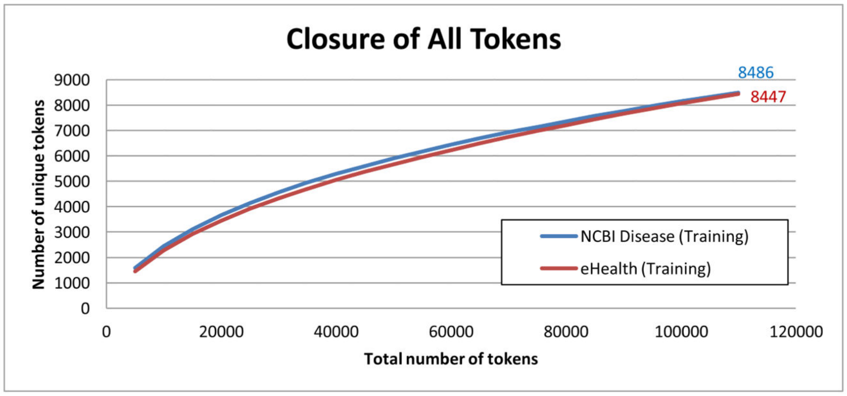


Figure 2.

A comparison of lexical closure curves between clinical narrative (eHealth) and biomedical publications (NCBI Disease) for the complete training corpora. These curves show that the trend toward closure is similar, implying that a difference in vocabulary richness is not the cause of the performance disparity.

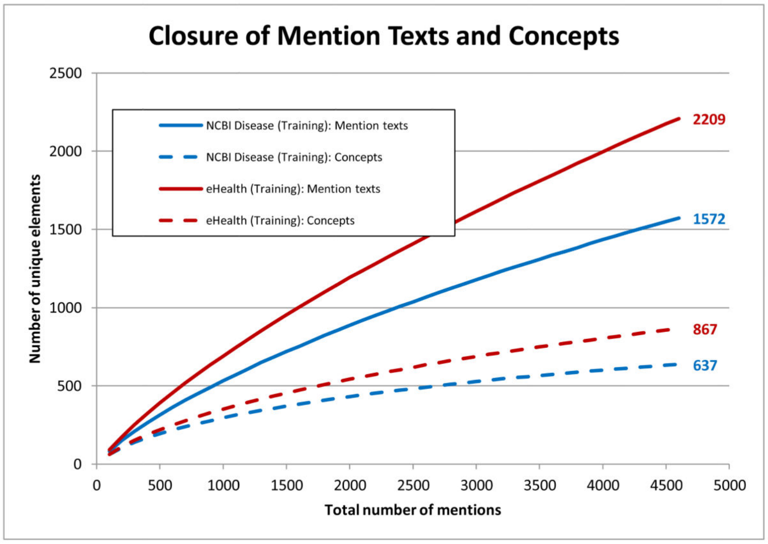


Figure 3.

Closure properties for mention texts (the full text of each mention as a single span) and concepts (the annotated concept identifier). Term variation is evident by the stronger trend toward closure for the concepts than for mention texts. Both mention texts and concepts show significantly less trend towards closure in the eHealth corpus, making disorder recognition more difficult and thus explaining the performance difference.

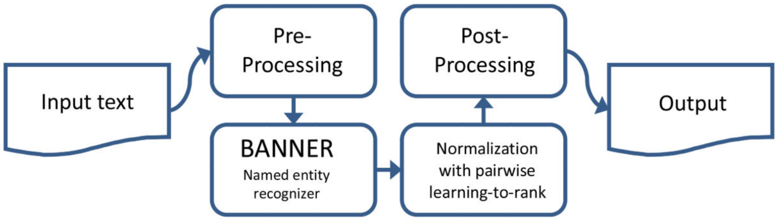


Figure 4.

Diagram of system components and processing flow

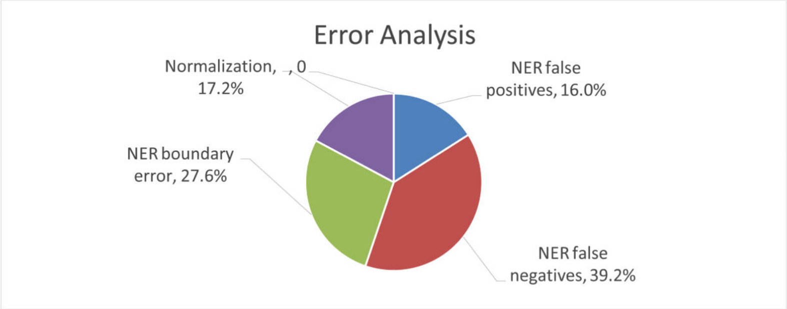


Figure 5.

Summary of the error analysis, illustrating that named entity recognition (NER) errors outnumber normalization errors by more than 4-to-1.

Table 1

Illustrative examples of common challenges in processing text from clinical narratives.

Category	Detail	Example
Flexible formatting	Variable formatting semantics	• **Section header:** “Admitting Diagnosis: SPLENOMEGALIA” • **Inseparable phrase:** “Neuro: nonfocal”
	Structure without sentences	• “Height: (in) 75 Weight (lb): 245 BSA (m2): 2.39 m2 BP (mm Hg): 92/52 HR (bpm): 120” • “Digoxin 250 mcg Tablet Sig: One (1) Tablet PO DAILY (Daily).”
	Missing punctuation	• **Commas:** “no m/r/g b/l coarse rhonchi w/ assoc upper airway sounds.” • **Periods:** “Well appearing male in no acute distress Chest is clear No hernias”
	Parenthetical expressions	• “Severe AS (AoVA <0.8cm2).” • “Mild (1+) aortic regurgitation is seen.”
Atypical grammar	Missing expected words	• **Verb:** “No pneumothorax.” • **Object:** “Lopressor beta blockade was given.” • **Articles:** “Ultrasound showed no evidence of [a] pseudoaneurysm.”
	Unusual part-of-speech combinations	• **Adjective without noun modified:** “Head, eyes, ears, nose, and throat examination revealed normocephalic and atraumatic.”
Rich descriptions	Variety of textual subjects	• **Patient:** “Awake, alert and oriented times three.” • **Anatomy:** “The left atrium is mildly dilated.” • **Test or procedure:** “Suboptimal image quality - poor echo windows.” • **Family:** “mother had CABG grandmother had CAD”
	Variety of communication styles	• **Diagnosis:** “He was also tachycardic” • **Evidence:** “He had a fingerstick of 142.”
	Language specific to medical context	• **Jargon:** “septic picture likely aspiration pneumonia secondary to dobhoff placement.” • **Ad-hoc abbreviations:** “His iron studies, LDH/Bili/Hapto, B12/Folate were normal.” • **Acronyms:** “recent tension PTX at OSH”
	Misspellings	• “Mitral stenosis is not present and definate [sic] mitral regurgitation is not seen.” • “s/p gsw now with fevers r/o abcess [sic]” • “Sclarea [sic] anicteric.” • “ventilator dependent respiratoy [sic] failure”

Table 2

Count and average size of each type of clinical note in the Corpus.

Note Type	Number in Training Set	Number in Test Set	Average Size (characters)
Discharge summary	61 (30.7%)	76 (76.0%)	7349
ECG report	54 (27.1%)	0 (0.0%)	285
Echo report	42 (21.1%)	12 (12.0%)	2237
Radiology report	42 (21.1%)	12 (12.0%)	1891

Table 3

Illustrative example of sentences and disorder annotations.

Example	Sentence	Annotation(s): CUI (Preferred name)
Exact match	No chest pain.	C0008031 ("Chest Pain")
Term variation	He should return to the ED immediately if any rash occurs	C0015230 ("Exanthema")
Partial term variation	Chief Complaint: left-sided facial droop	C0427055 ("Facial Paresis")
Abbreviation	The patient was found to have left lower extremity DVT.	C0340708 ("Deep vein thrombosis of lower limb")
Morphology	Pertinent physical findings reveal that her sclerae were anicteric.	CUI-less ("not icteric")
CUI-less	The patient was admitted with low blood pressure.	CUI-less ("hypotension")
Disjoint mention	A tumor was found in the left ovary.	C0919267 ("ovarian neoplasm")
Inference	On admission, blood glucose was 705.	*<none>*

Table 4

Examples of IO and CD labels in a sentence with one contiguous mention (“hiatal hernia”) and one disjoint mention (“laceration | esophagus”). Unlike the IO labeling, the CD labeling differentiates between contiguous mentions and spans that are part of a disjoint mention, allowing the latter to be handled separately.

Text	**EGD**	**showed**	**hiatal**	**Hernia**	**and**	**laceration**	**in**	**distal**	**esophagus**
IO labels	O	O	I	I	O	I	O	O	I
CD labels	O	O	C	C	O	D	O	O	D

Table 5

Table of results on the eHealth Test set. The highest value for each measure is shown in bold. The improvement between the base version (first row) and final version (last row) are statistically significant for all measures (p < 0.01; bootstrap resampling [43]).

	Span-only		**Span + Concept**	
Version	**Strict Precision/Recall/F-score**	**Relaxed Precision/Recall/F-score**	**Strict Precision/Recall/F-score**	**Relaxed Precision/Recall/F-score**
Base version	0.773/0.664/0.714	0.917/0.795/0.852	0.688/0.591/0.636	0.745/0.641/0.689
Base version + improved dictionary features	0.788/0.675/0.723	0.917/0.801/0.855	0.693/0.602/0.644	0.746/0.649/0.694
Base version + improved dictionary features & MetaMap feature	0.794/0.700/0.744	0.922/0.819/0.867	0.708/0.624/0.663	0.757/0.668/0.710
Base version + improved dictionary features, MetaMap feature & post-processing	**0.797/0.713/0.753**	**0.923/0.831/0.875**	**0.712/0.637/0.672**	**0.760/0.681/0.719**

Table 6

Results comparison for highest-performing version of system, using n-best decoding in the conditional random field NER model. This increases recall at the expense of precision.

	Span-only		**Span + Concept**	
Version	**Strict Precision/Recall/F-score**	**Relaxed Precision/Recall/F-score**	**Strict Precision/Recall/F-score**	**Relaxed Precision/Recall/F-score**
Best version	**0.797**/0.713/**0.753**	**0.923**/0.831/**0.875**	**0.712**/0.637/**0.672**	**0.760**/0.681/**0.719**
N-best, n=5	0.698/0.759/0.728	0.883/0.872/0.868	0.619/0.673/0.645	0.678/0.716/0.696
N-best, n=50	0.091/**0.889**/0.165	0.118/**0.974**/0.138	0.073/**0.744**/0.138	0.088/**0.805**/0.159