Modern Clinical Text Mining: A Guide and Review

Bethany Percha¹,²

¹Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA, 10025; email: bethany.percha@mssm.edu
²Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA, 10025

Keywords
text mining, natural language processing, electronic health record, clinical text, machine learning

Abstract

Electronic health records (EHRs) are becoming a vital source of data for healthcare quality improvement, research, and operations. However, much of the most valuable information contained in EHRs remains buried in unstructured text. The field of clinical text mining has advanced rapidly in recent years, transitioning from rule-based approaches to machine learning and, more recently, deep learning. With new methods come new challenges, however, especially for those new to the field. This review provides an overview of clinical text mining for those who are encountering it for the first time (e.g., physician researchers, operational analytics teams, machine learning scientists from other domains). While not a comprehensive survey, it describes the state of the art, with a particular focus on new tasks and methods developed over the past few years. It also identifies key barriers between these remarkable technical advances and the practical realities of implementation at health systems and in industry.
1. INTRODUCTION

Among the most significant barriers to large-scale deployment of electronic health records (EHRs) in quality improvement, operations, and research is the amount of EHR data stored as unstructured text (1). Structured, machine computable data, such as procedure and diagnosis codes, are in the minority. The bulk of information relating clinical findings to decisions, and communicating the logical and deductive processes of medicine, is buried within progress notes, radiology and pathology reports, and other free text documents (2, 3). Examples include:

- Treatment goals and outcomes (e.g. success or failure of treatments, criteria for success, decisions about subsequent treatments)
- Interpretations of radiology and pathology images and laboratory test results
- Social determinants of health (e.g. social connection/isolation, housing issues, mentions of financial resource strain) (4)
- Symptoms, symptom changes, and their interpretation (5)
- Past medical history and family history
- Patient’s emotional disposition, mood, and interactions with health providers
- Detailed descriptions of procedures (e.g. labor and delivery, heart catheterization, imaging studies, surgeries)
- Adherence to treatment plans (e.g. medications, physical therapy, procedures)
- Allergies, side effects, and other adverse events
- Results of physical examination (e.g. review of systems and interpretation of findings)
- Patient’s reasons for seeing a health provider; primary and secondary complaints
- Psychiatric evaluations and records of therapy sessions
Discharge summaries and follow-up plans

Some have speculated that modern machine learning algorithms, combined with EHR and other patient data, will enable the convergence of human and machine intelligence in healthcare (6, 7). From a practical standpoint, such a vision hinges on text mining. Without the ability to reliably process and interpret vast quantities of clinical text, all attempts to create high-performance predictive models, phenotyping algorithms, and data-driven treatment strategies (“precision medicine”) will face substantial challenges.

For the past several decades, a community of researchers working at the intersection of computer science and medicine has developed strategies for information extraction and modeling of clinical text, using techniques somewhat distinct from those of the broader natural language processing (NLP) research community (8, 9). Their efforts have led to the development of new methods and the production of both commercial (10) and open-source (11) software systems for clinical text mining. In recent years, technology giants like Amazon and Google have also recognized the importance of clinical text mining and joined the fray; Amazon Comprehend Medical (12) now comes packaged as a software add-on to Amazon Web Services, incentivizing storage of EHR data on Amazon’s HIPAA-compliant cloud platform by providing seamless clinical text processing. Dedicated clinical text processing companies such as (as of this writing) Clinithink (clinithink.com), Linguamatics (linguamatics.com), and Apixio (apixio.com) have built proprietary systems of their own, promising to improve clinical trial recruitment, disease registry creation, government reporting, and billing, all through improved mining of unstructured clinical text.

As a data scientist with a background in biomedical text mining, I am frequently approached by physician colleagues and academic and industry collaborators who, for various reasons, have found themselves needing to process clinical text. Many perceive clinical text mining to be a “solved” problem, believing that one can simply apply a packaged clinical NLP system to extract structured data for a variety of downstream applications. As a result, I often find myself explaining the limits of current NLP technology and the fact that clinical NLP encompasses many different goals, progress on some of which is further along than others. The purpose of this review, therefore, is to provide a starting point for those who are encountering clinical text mining for the first time. Far from a comprehensive survey, it focuses on a subset of methods and ideas that are particularly clear and generalizable and can serve as starting points for further explorations of the field. Importantly, nothing I discuss here requires access to institution-specific or proprietary software, rule sets, or training corpora. My goal is to provide “outsiders” with a realistic baseline for what it is possible to accomplish with clinical text mining today.

2. A SHORT TAXONOMY OF TASKS AND APPROACHES

2.1. Information Extraction vs. Modeling

Clinical text can play multiple roles in a project, so it is important to start by defining one’s overall goal and how the text fits in. For example, electronic phenotyping algorithms (13, 14, 15, 16) often combine clinical notes with structured data, such as diagnosis codes, medication orders, and procedures, to make a prediction about whether a patient has a disease or other phenotype. Here the primary goal of text mining is information extraction: converting the text into a set of structured features that can be combined with other types of information.
of features to produce an answer (17). EHR search indexing, knowledge base construction, and patient timeline building are similar in their focus on information extraction.

Equally important, however, are problems where the goal is to make a prediction or inference from the text itself – for example, to classify mammography reports by BI-RADS category (18) or to cluster clinical documents to uncover latent structure (19). This may or may not require a separate information extraction step. For example, methods such as end-to-end, deep learning-based text classification models (20, 21, 22), which produce answers directly from the raw text, often shine in such cases. One important consideration is whether human-interpretable features are necessary for the project or whether the algorithm can be allowed to learn its own representations of text automatically in the course of solving a downstream task (22).

2.2. Rule-Based vs. Statistical Approaches

Clinical NLP systems fall into two broad categories: rule-based and statistical. Rule-based systems codify expert knowledge into a set of structured rules, or templates, which produce structured information when applied to unstructured text. For example, a rule might specify patterns of words, phrases, or parts of speech that signal the presence of a particular type of entity; e.g. “if the word ‘received’ is followed by a noun followed by ‘for’ and then a disease name, assume the noun is a drug name”. Many of the best-performing clinical NLP systems are rule-based: of 263 clinical text mining articles reviewed by Wang et al in 2018 (17), 171 (65%) used rule-based methods. However, rule-based systems have two important disadvantages. First, domain experts must often expend substantial time and effort to construct the rules. Second, because they are domain-specific, they do not generalize well to new problems; a rule-based system for identifying drug names in text will not be good at anything other than identifying drug names in text.

The alternative is a system built using a statistical learning (“machine learning”) algorithm. If provided with some text in which all of the drug names are labeled, for example, the algorithm will try to identify patterns that indicate a particular span of text is a drug name (9, Ch. 8). Learning algorithms themselves are often task-independent, which is one of their key advantages. However, statistical learning algorithms require annotated training data, which in the clinical domain is often limited or nonexistent (23). In addition, privacy concerns often make it impossible to share training data across institutions. As a result, while the NLP community has increasingly turned toward machine learning and away from rules-based approaches, clinical text mining maintains a strong focus on rules (8).

3. SOFTWARE FOR CLINICAL INFORMATION EXTRACTION

The three most common information extraction tasks – named entity recognition (37, 38, 39), concept normalization (40, 41), and relation extraction (Section 7) – are still active areas of research. However, because of the broad need for basic information extraction in applied tasks like medical coding, search, and case finding, software systems have been developed to perform these tasks automatically. This section reviews current state-of-the-art methods and systems and provides examples of the type of output one can expect from each system.
CLINICAL TEXT MINING SOFTWARE & RESOURCES

The following tools are popular choices for general and clinical text processing (e.g. word and sentence tokenization, part-of-speech tagging, chunking, parsing, NER, word and phrase embeddings). The first section contains general-purpose libraries, while the second contains resources specific to clinical text.

<table>
<thead>
<tr>
<th>Resource</th>
<th>Language</th>
<th>URL</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLTK Toolkit</td>
<td>Python</td>
<td>nltk.org</td>
<td>(24)</td>
</tr>
<tr>
<td>Stanford CoreNLP</td>
<td>Java</td>
<td>stanfordnlp.github.io/CoreNLP</td>
<td>(25)</td>
</tr>
<tr>
<td>Stanza</td>
<td>Python</td>
<td>stanfordnlp.github.io/stanza</td>
<td>(26)</td>
</tr>
<tr>
<td>spaCy</td>
<td>Python, Cython</td>
<td>spacy.io</td>
<td>(27)</td>
</tr>
<tr>
<td>scispacy</td>
<td>Python</td>
<td>allennlp.github.io/scispacy</td>
<td></td>
</tr>
<tr>
<td>Apache OpenNLP</td>
<td>Java</td>
<td>opennlp.apache.org</td>
<td></td>
</tr>
<tr>
<td>CRFSuite</td>
<td>Python</td>
<td>chokkan.org/software/crfsuite</td>
<td></td>
</tr>
<tr>
<td>Scikit-learn</td>
<td>Python</td>
<td>scikit-learn.org</td>
<td></td>
</tr>
<tr>
<td>Gensim</td>
<td>Python</td>
<td>radimrehurek.com/gensim/index.html</td>
<td>(28)</td>
</tr>
<tr>
<td>BERT</td>
<td>Python</td>
<td>github.com/google-research/bert</td>
<td>(29)</td>
</tr>
<tr>
<td>MetaMap</td>
<td>Java</td>
<td>metamap.nlm.nih.gov</td>
<td>(30)</td>
</tr>
<tr>
<td>MetaMap Lite</td>
<td>Java</td>
<td>metamap.nlm.nih.gov/MetaMapLite.shtml</td>
<td>(31)</td>
</tr>
<tr>
<td>cTAKES</td>
<td>Java</td>
<td>ctakes.apache.org</td>
<td>(11)</td>
</tr>
<tr>
<td>Stanza clinical</td>
<td>Python</td>
<td>stanza.run/bio</td>
<td>(32)</td>
</tr>
<tr>
<td>DNORM</td>
<td>Java, REST API</td>
<td>ncbi.nlm.nih.gov/research/bionlp/Tools/dnorm</td>
<td>(33)</td>
</tr>
<tr>
<td>Clinical BERT</td>
<td>Python</td>
<td>github.com/EmilyAlsentzer/clinicalBERT</td>
<td>(34)</td>
</tr>
<tr>
<td>UMLS</td>
<td>N/A (extraction software in Java)</td>
<td>nlm.nih.gov/research/umls/index.html</td>
<td>(36)</td>
</tr>
</tbody>
</table>

3.1. Named Entity Recognition

Named entity recognition (NER) is the task of identifying and locating mentions of conceptual categories, such as drug, symptom, or disease names, in text. It is perhaps the most widely-studied information extraction task, and existing clinical NER systems can already identify a variety of clinically-relevant entities, including problems, tests, and treatments (32, 37, 38), medication and adverse event names (42, 43), and protected health information (PHI) (44, 45). Figure 1 shows the raw output from Stanza (26, 32), a state-of-the-art NER system trained to tag clinical entities using test, problem, and treatment concept annotations from the 2010 i2b2/VA dataset (46).

The simplest NER systems are dictionary-based: they simply compare text strings to a list of terms from a specific category, such as disease names. While these approaches are common and frequently yield acceptable performance on clinical text (47, 48), modern clinical NER systems more commonly employ machine learning models adapted for sequence data, including conditional random fields (CRFs), recurrent neural networks (RNNs), and RNN variants such as long short-term memory networks (LSTMs). For example, Stanza (Figure 1) uses pretrained character-level language models (49) fed into Bi-LSTM-CRF sequence taggers (32, 50, 51). Trained using corpora hand-annotated with entity type(s)
Progress Note:

Ms. S. is a 43F h/o antiphospholipid syndrome, HTN, DM here for routine follow-up appointment.

Her main concern today is a 3 month h/o worsening shortness of breath a/w lower extremity edema. Her exercise tolerance has decreased from 10 blocks to half a block over the past few years.

She denies chest pain, palpitations, orthopnea, calf tenderness, fevers, and substance use.

On exam, she is afebrile, BP 110/80, HR 80, RR 18, O2 Sat 98% on room air. She was A0x3, +JVD to her mid-neck, +HJ reflux. Normal st, prominent P2. Lungs CTAB, no wheezes, rales or rhonchi. Abdomen soft, non-tender, non-distended, +bowel sounds. 1+ lower extremity pitting edema bilaterally to the shins.


Chest X ray was clear.

EKG with a HR 84, NSR, no axis deviation, no ischemic changes.

Assessment:

Ms. S. is a 43F h/o antiphospholipid syndrome, HTN, DM with progressive dyspnea and lower extremity edema concerning for new-onset acute decompensated heart failure. Given that she is afebrile without any infectious symptoms, pneumonia is less likely. She does not smoke which makes COPD exacerbation less likely.

Plan:

- BNP
- Transthoracic echocardiogram

Figure 1

A sample clinical progress note (not a real patient) with named entity annotations provided by the Stanza clinical text processing pipeline, trained using data from the 2010 i2b2/VA challenge. The Stanza pipeline tags three types of named entities: treatment, problem, and test. For this particular note, no treatment entities were found. Medical terms, abbreviations and acronyms: HTN: hypertension; DM: diabetes mellitus; h/o: history of; a/w: along with; A0x3: alert and oriented to person, time, and place; JVD: jugular vein distention; HJ reflux: hepatojugular reflux, distention of jugular vein produced by applying manual pressure to the liver; RRR: regular rate and rhythm (of pulse); S1: heart sound produced by closure of atrioventricular (mitral and tricuspid) valves; P2: heart sound produced by closure of pulmonic valve; CTAB: “clear to auscultation bilaterally”, an abbreviation used in lung examinations; pitting edema: if area of swelling pressed, a pit remains; Hgb: hemoglobin, measured in units of g/dL; Cr: creatinine, measured in units of mg/dL; NSR: normal sinus rhythm; BNP: brain natriuretic peptide test (indicative of heart failure).
of interest, these algorithms learn to identify features of a text string and its surrounding context that predict whether it is one of the desired types.

What constitutes a “feature” depends on the system. Traditionally, NER algorithms have selected features from predefined sets, including morphological (capitalization and punctuation patterns, presence/absence/location of numbers, etc.), syntactic (parts of speech, grammatical dependencies, etc.), semantic (membership in a lexicon, position in an ontology, etc.), and other specialized or hand-coded features (52). Some systems incorporate pre-trained word or phrase embeddings (Section 4), and modern neural network-based NER systems often learn higher-order embeddings directly from patterns in the text itself (50, 53).

A major advantage of machine learning-based NER systems is their flexibility. The same system can often learn to tag different types of entities simply by swapping training datasets. For example, while Figure 1 shows “test”, “problem”, and “treatment” annotations, the Stanza system has also been trained using a corpus of 150 chest CT radiology reports (54) to tag “anatomy”, “anatomy modifier”, “observation”, “observation modifier”, and “uncertainty” concepts (32). The reverse is also true: different machine learning algorithms can be trained using the same training data. In fact, when looking for an NER system or any other clinical NLP system, a useful strategy is to identify an annotated corpus for that task and look for papers that have cited the corpus. For example, like Stanza, the Clinical Named Entity Recognition ( CliNER ) system (38) was trained using concept annotations from the 2010 i2b2/VA NLP challenge (46). Other widely cited systems trained on the same dataset are the Bi-LSTM-CRF systems by Chalapathy et al (55) and Unanue et al (56) and Tang et al’s system combining support vector machines (SVMs) with CRFs (57).

3.2. Domain Specificity and Key Challenges

It is worth pausing here to consider the conceptual and practical challenges illustrated by Figure 1. First, it is clear that NER only makes sense when the entities involved are discrete and have well-defined locations in text. Many clinically important concepts, such as income, housing, and employment history, are unlikely to be described using simple and consistent terminology that can be picked up by NER algorithms. Second, not all of the NER annotations in Figure 1 are correct or meaningful without consideration of the surrounding context. Labeling the term “acute decompensated heart failure” as a problem, for example, means little without the qualification that it is a suspected, not definite, diagnosis.

There are also practical concerns regarding NER model training and maintenance. Although dozens of different clinical NER systems have been developed, many are now obsolete, and not all are released as “production-ready” code (i.e. easy to download and use). In addition, if one is interested in an entity class for which no pre-annotated corpus or pre-trained model is available, there is no alternative but to train one’s own system; this means either defining a rules-based approach or creating a custom, annotated training set.

Finally, there is the issue of domain specificity. Clinical text is complex, incorporating specialized medical terms, numerical measures and scores, abbreviations (see Figure 1 caption), misspelled words, and poor grammar (20). Many high-quality general domain NER models exist, such as those from the Stanford CoreNLP (25) and spaCy libraries (https://spacy.io). However, these are trained using general domain text, such as telephone conversations, newswire, newsgroups, broadcast news, broadcast conversation, and
Table 1  Examples of cTAKES annotations associated with the note in Figure 2.
The annotations in the top section are correct mappings, and those in the bottom section
are incorrect mappings. There were 122 unique cTAKES annotations for this note.

<table>
<thead>
<tr>
<th>Line Number</th>
<th>Annotation Type</th>
<th>Original String</th>
<th>Normalized Term</th>
<th>UMLS Concept ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>DiseaseDisorderMention</td>
<td>HTN</td>
<td>Hypertensive disease</td>
<td>C0020538</td>
</tr>
<tr>
<td>4-5</td>
<td>SignSymptomMention</td>
<td>shortness of breath</td>
<td>Dyspnea</td>
<td>C0013404</td>
</tr>
<tr>
<td>7</td>
<td>SignSymptomMention</td>
<td>chest pain (negated)</td>
<td>Chest pain (negated)</td>
<td>C0008031</td>
</tr>
<tr>
<td>10</td>
<td>SignSymptomMention</td>
<td>JVD</td>
<td>Jugular venous engorgement</td>
<td>C0425687</td>
</tr>
<tr>
<td>16</td>
<td>ProcedureMention</td>
<td>EKG</td>
<td>Electrocardiography</td>
<td>C1623258</td>
</tr>
<tr>
<td>24</td>
<td>ProcedureMention</td>
<td>Transthoracic echocardiogram</td>
<td>Transthoracic echocardiography</td>
<td>C0430462</td>
</tr>
<tr>
<td>10</td>
<td>DiseaseDisorderMention</td>
<td>reflux</td>
<td>Gastroesophageal reflux disease</td>
<td>C0017168</td>
</tr>
<tr>
<td>11</td>
<td>MedicationMention</td>
<td>CTAB</td>
<td>Cetrimonium bromide</td>
<td>C0951233</td>
</tr>
<tr>
<td>23</td>
<td>MedicationMention</td>
<td>BNP</td>
<td>Nesiritide</td>
<td>C0054015</td>
</tr>
</tbody>
</table>

blogs. As such, they tag somewhat generic entities like person, number, and place names,
which may or may not be relevant in a clinical text mining context. A second class of sys-
tems are those that have been trained using biomedical text, usually from PubMed research
articles and abstracts. Recent examples are the scispaCy library (27) and the transformer-
based language model BioBERT, fine-tuned for NER (58). These systems often tag entity
types that are relevant to clinical text, such as gene names; however, because they were
trained using scientific writing, they may suffer reduced accuracy on clinical text. The
issue of domain specificity and the need for domain-specific models extends beyond NER,
affecting nearly all tasks in clinical text mining.

3.3. Concept Normalization

The output of a clinical NER system (Figure 1) is a set of named entities of one or more
types. The obvious downside to such output is that it tells one nothing about the entities
except their type(s); for example, there is no way of knowing that the strings “HTN” and
“hypertension” – even if they are in the same note and both labeled as problems – refer to
the same concept. Likewise, although an NER system may recognize multi-word phrases
(e.g. “lower extremity pitting edema bilaterally to the shins”, Line 13, Figure 1), it does
not understand how the component words contribute to the meaning of each phrase, and
it cannot easily connect a given phrase to coreferent phrases, even from the same passage
(e.g. “lower extremity edema”, Line 5, Figure 1).

Concept normalization, also known as entity linking, is the task of assigning a unique
identity to each entity name mentioned in text. In the clinical domain, this typically involves
mapping each entity name to a known concept from a structured terminology or ontology.

Concept Normalization: The task of assigning a unique identity to an entity name
recognized in the text. In the biomedical domain, this typically involves mapping
the name to a known concept from a structured terminology or ontology.
Ms. S. is a 43F h/o antiphospholipid syndrome, HTN, DM, here for routine follow-up appointment.

Her main concern today is a 3 month h/o worsening shortness of breath a/w lower extremity edema. Her exercise tolerance has decreased from 10 blocks to half a block over the past few years.

She denies chest pain, palpitations, orthopnea, calf tenderness, fevers, and substance use.

On exam, she is afebrile, BP 110/80, HR 80, RR 18, O2 Sat 98% on room air. She was AOx3, +JVD to her mid-neck, +HJ reflux. RRR, normal s1, prominent P2. Lungs CTAB, no wheezes, rales, or rhonchi. Abdomen soft, non-tender, non-distended, +bowel sounds. 1+ lower extremity pitting edema bilaterally to the shins.

Labs notable for Hgb 12, platelet count of 350. Na 135, K 3.5, Cr 1. Chest X ray was clear.

EKG with a HR 84, NSR, no axis deviation, no ischemic changes.

Assessment:
Ms. S. is a 43F h/o antiphospholipid syndrome, HTN, DM, with progressive dyspnea and lower extremity edema concerning for new-onset acute decompensated heart failure. Given that she is afebrile without any infectious symptoms, pneumonia is less likely. She does not smoke which makes COPD exacerbation less likely.

Plan:
- BNP
- Transthoracic echocardiogram

The same clinical progress note as in Figure 1, with annotations provided by the cTAKES (version 4.0) default pipeline. The abbreviations are the same as in Figure 1. The cTAKES pipeline detects negation and uncertainty and maps each entity to its corresponding concept in UMLS. A selection of the UMLS concepts found in this note is in Table 1.
building and deploying clinical NLP pipelines, including UMLS mapping.

The same note analyzed by Stanza in Figure 1 is shown in Figure 2, this time with annotations produced by cTAKES, a popular system developed at the Mayo Clinic (11). A selection of the 122 detailed UMLS mappings produced by cTAKES is in Table 1. In addition to UMLS-based concept normalization, cTAKES detects negation (63), uncertainty, and experiencer (whether the statement refers to the patient or, e.g. a family member). The results shown in Figure 2 are from the default cTAKES pipeline, i.e. what one could expect running cTAKES “out of the box”. Most of the annotations are correct; for example, cTAKES correctly maps the string “HTN” to the normalized concept Hypertensive disease (CUI C0020538) and understands that “shortness of breath” is a synonym for Dyspnea (CUI C0013404). A key shortcoming, however, is cTAKES’ reliance on dictionary-based lookups to identify and normalize named entities. This is apparent in Figure 2, where cTAKES labels the strings “CTAB”, “BNP”, and “Hgb” as medications because of spurious UMLS mappings (e.g. “CTAB” maps to “cetrimonium bromide” in UMLS). If the specificity of extracted medication terms were crucial for one’s application, therefore, it might make sense to include a dedicated named entity recognition system for medication names in the cTAKES pipeline. In addition, depending on the application, full concept normalization may not be necessary; in one recent study (64), using cTAKES annotations as features in a 30-day readmission model yielded no better performance than N-grams.

Like named entity recognition, clinical concept normalization is still an active area of research. For those interested in this task, a good place to start are the disorder normalization systems built for the SHARE/CLEF eHealth 2013 Evaluation Lab, a community NLP challenge focusing on clinical named entity recognition and concept normalization (65, 41). DNorm (33, 61) was the top-performing system on the concept normalization task, deploying a pairwise learning-to-rank approach that was the first of its kind in the clinical concept normalization literature. More recent studies have applied deep learning models to the same task and dataset (66, 67).

3.4. Numbers, Ranges, and Sections

There are a few information extraction tasks of particular importance to clinical text for which dedicated systems have been developed. These systems are generally rule-based and rely on regular expressions (68). For example, extraction of lab values and vital signs is a distinct task from named entity recognition because it requires interpreting numeric values and ranges. The Valx system (69) extracts and structures lab test comparison statements, though so far it has only been applied to trial descriptions from ClinicalTrials.gov. The CNN-based system developed by Xie et al (70) identifies blood pressure readings, determines the exactness of the readings, and classifies the readings into three classes: general, treatment, and suggestion. Their machine learning-based workflow could be adapted to extract other types of numeric values.

Section identification is another task somewhat unique to the clinical text mining literature. It involves identifying the section labels associated with each span of text within a note (e.g. Progress Note, Assessment, and Plan in Figure 1), which informs the interpretation of whatever is found there. To date, the only section identification system used outside the institution in which it was developed is the SecTag system by Denny et al (71). A complete review of section identification methods and systems can be found in (72).
4. EMBEDDINGS AND PRETRAINING

The core idea behind concept normalization (Section 3.3) is semantic relatedness; two terms can look different, yet refer to the same concept. However, semantic relatedness extends beyond the dichotomy of same vs. different; terms can have degrees of similarity (e.g. “dog” vs. “cat” as opposed to “dog” vs. “volcano”) and can be similar in different ways (e.g. “queen” vs. “king” as opposed to “queen” vs. “president”). Modern NLP systems represent this idea mathematically using a construct called an embedding.

4.1. Word, Phrase, and Character Embeddings

An embedding is a semantically-meaningful mathematical representation of a word, phrase, or other piece of text. Usually a vector, it is designed in such a way that words and phrases with similar meaning have similar vectors. Meaning is difficult to represent using numbers, so embedding methods replace “meaning” with “context” and build vectors to reflect usage patterns, typically within large, unlabeled corpora. The NLP subfield of distributional semantics, which originated with Latent Semantic Analysis in 1988 and reached a milestone with the development of word2vec (73) and GloVe (74) in 2013–2014, is a collection of methods all built around the central goal of creating vector-space embeddings of words and phrases that reflect how they are used in context. To compare the meaning of two words, one simply calculates the cosine similarity of their corresponding vectors.

From a clinical text mining standpoint, embeddings are useful in two ways. First, because they do not require annotated corpora for training, it is easy to create embeddings that are specific to clinical text, or that capture regularities of expression within a particular clinical subfield or institution. These will often outperform general-domain embeddings on clinical text mining tasks (53). Specialized clinical text embeddings have been used to improve clinical NER (75), resolve abbreviations in clinical text (76), expand a structured lexicon of radiology terms (77) and build specialized lexicons from scratch (78). Second, an embedding can incorporate structured information beyond what is found in the text (79), and embeddings have been created to represent CUIs (80), documents (81, 82), or entire patient records (83). Any task in which the notion of similarity is important, particularly when that similarity is based on patterns in text, can probably benefit from embeddings.

For more information about embeddings, readers are encouraged to consult Turney and Pantel (84) for a review of early methods and Kalyan et al (85) for a review of embedding methods currently in use in clinical text mining.

4.2. Contextual Embeddings and Pretraining

Until the last few years, embeddings consisted of one vector per entity; that is, one vector per word, phrase, or document. However, novel neural network architectures (22) have permitted the creation of embeddings that vary depending on the context; this has expanded the representational power of embedding methods and led to the creation of massive pretrained language models like BERT (Bidirectional Encoder Representations from Transformers) (29) and GPT-3 (openai.com). These models are generally too resource-intensive to be trained from scratch. Instead, a transfer learning approach (86) is used in which models trained on general-domain corpora are either further pre-trained or fine-tuned on clinical text for use in clinical text mining tasks (85). For example, Alsentzer et al recently trained BERT models on 2 million notes from the MIMIC-III (87) database. They produced two...
models, one for generic clinical text and another for discharge summaries, which they released publicly (34). They and others have demonstrated that BERT models fine-tuned on clinical corpora improve the state of the art on clinical NER, de-identification, inference, and concept normalization tasks (88, 89), though in at least one case, UMLS features still contributed valuable additional information (90).

The downside of these models is that they require some technical sophistication to adapt and apply. Whereas the original word2vec could be run on a plain text corpus using a single script and output vectors to a text file, to use BERT requires knowledge of how to “wire up” a pre-trained model to task-specific output layers for fine-tuning. However, it is likely that end-to-end clinical text processing systems, like cTAKES, will begin to incorporate BERT and related methods into different annotation modules as the technology develops.

5. TEXT CLASSIFICATION

Text classification is perhaps the most sought-after application of clinical text mining. A recent survey (22) found that of 212 clinical text mining papers employing deep learning methods, 88 (41.5%) focused on text classification; text classification and NER together encompassed 75.5% of articles. The goal of text classification is to classify documents (or sentences, phrases, etc.) into two or more discrete categories. Examples from the clinical domain include classifying primary care descriptions of back pain into acute vs. lower back pain (91), distinguishing normal vs. abnormal knee MRI reports (92), and assessing whether a patient is a current or former smoker vs. a non-smoker based on clinical notes (93). Text classification is a modeling task—typically, it is its own goal. Often it will incorporate features identified through information extraction (Section 3), like named entities or CUIs, or embeddings (Section 4).

A recent systematic review of clinical text classification describes standard text classification algorithms, as well as popular approaches to preprocessing, feature selection, and training set construction (20). An older but still relevant review surveys text classification methods for automated clinical coding (94). In general, text classification methods for clinical text are similar to those for other domains, with the exception that specialized medical resources, such as UMLS, often serve as additional sources of features.

5.1. Feature Construction and Selection

The use of individual words or N-grams as features, while common in text classification (95, Ch. 13), often results in undesirable levels of feature sparsity when applied to clinical text. As a result, feature selection and dimensionality reduction methods are of particular importance in clinical text classification. Feature selection based on TFIDF weighting (95, Ch. 6) is common, as are embeddings (Section 4), which turn a potentially unmanageable number of word and text features into dense representations of fixed dimensionality (96). Concept normalization (Section 3.3) also plays a particularly important role in clinical text classification; it is common to preprocess clinical text with a system like cTAKES or MetaMap to merge different term and phrase variants into the same structured concept, then use those concepts in a classification model (97, 98). It is also possible to exploit parent-child relationships from the UMLS hierarchy to create additional features, e.g. by including all parent terms for a given concept. Such ontology-guided feature engineering has been shown to improve performance on downstream clinical text classification tasks (99).
Finally, one can choose a classification algorithm that provides implicit feature selection. In one study, elastic net (100) was used to classify ICU patients into risk strata based on the text of nursing notes. It reduced the number of text features by over a thousandfold while maintaining near-optimal performance (101).

5.2. Deep Learning for Clinical Text Classification

Aside from those that have employed task-specific rules (Section 2.2), the majority of clinical text classification studies to date have used standard supervised machine learning algorithms, including support vector machines, naive Bayes, random forests, and boosting (92, 102, 103). However, over the past five years, deep learning algorithms have begun to displace other classifiers. One of their key advantages is a reduced need for feature engineering; embeddings of words, phrases, and higher-order text structures can be learned as part of the overall training process or incorporated via transfer learning from other pre-trained models. Several studies have deployed convolutional neural networks (CNNs) with high success on a variety of clinical text classification tasks: assigning diagnosis codes (104, 105), classifying radiology reports (21, 106), subtyping diseases (91), and determining the presence or absence of comorbidities (107). Alternative neural network architectures, such as LSTMs and attention networks, are commonly used in text classification tasks in the general NLP domain, although as of this writing, CNNs have been the dominant architecture in clinical text classification (22, 108). One recent paper exemplifies the end-to-end deep learning approach to clinical text classification, tying rule-based features together with word and UMLS-based concept embeddings in a single CNN-based classifier (107).

6. WEAK AND DISTANT SUPERVISION

As discussed in Section 2.2, machine learning approaches to clinical NLP generally suffer from a lack of training data (23). In addition, existing clinical information extraction and text classification models have generally been trained using the same few annotated datasets (46, 109, 110, 65, 111), which restricts the range and quality of annotations they produce. Most applied clinical text mining projects will therefore confront, at some point, the problem of insufficient or inappropriate training data. Two practical solutions to this problem are weak and distant supervision. Weak supervision is the act of creating “silver standard” training data by applying a weak, or noisy, labeling function to large amounts of unlabeled data. Distant supervision is a related practice in which external data sources, such as knowledge bases, are used as training signals. One can, in fact, view distant supervision as a form of weak supervision, and in practice the terms are often used interchangeably.

The paradigmatic clinical text mining example of distant supervision is using structured information from the EHR, such as ICD codes, as a labeling mechanism for unstructured text documents. For example, outcomes such as in-hospital mortality (16), hospital readmission (112, 113), and reportable adverse events (114) are routinely captured in the course of health system operations. Although this information is typically attached to patients or encounters, not individual text documents, one can use it as a source of noisy training labels for discharge summaries or other narrative documents attached to the encounters. These noisy labels then serve as a source of supervision for text classification algorithms. Similar results have been achieved using structured ICD9/10 diagnosis (115, 116, 117) and procedure codes (118) as class labels. However, this technique is somewhat limited to the
task of document classification; to obtain labels for specific words or text spans (i.e. for NER or relation extraction), one needs a labeling mechanism that works directly on the text.

An alternative is to apply simple heuristic rules to create noisy labels. For example, Wang et al used keyword-based weak labels for two separate tasks: smoking status classification and hip fracture classification (93). Importantly, they noted that their best-performing deep learning classifier, a CNN, was robust to the massive label noise created by the weak labeling. Their paper was, to my knowledge, the first to apply a combination of weak supervision and deep learning to clinical text classification; most earlier applications of weak supervision in the biomedical domain focused on images or text from biomedical research articles. Two earlier studies of note in the biomedical domain are Sabbir et al’s study of distant supervision for biomedical word sense disambiguation (119) and Fries et al’s description of the SwellShark system (120), a generative model for biomedical NER that uses lexicons and ontologies for weak labeling. The Snorkel system, on which SwellShark is based, was recently used to weakly label clinical notes for the purposes of extracting implant details and reports of complications and pain after hip replacement; the weakly labeled notes were then used to train deep learning models to recognize (pain, anatomy) and (complication, implant) relations (121). These methods improved classification performance by 12.8-53.9% over rule-based methods and detected over six times as many complication events compared to structured data alone.

Alternative approaches to the efficient annotation of training sets for clinical text mining include crowdsourcing and active learning. Crowdsourcing is not usually a viable option in the clinical domain because of privacy concerns. Active learning is a strategy for minimizing annotation effort by iteratively sampling subsets of data for human annotation based on the current performance of a supervised learning algorithm (122, 123). However, it still requires recruiting one or more experts to create the annotations.

7. RELATION EXTRACTION AND INFERENCE

Relation extraction is the task of assigning a structured form to a relationship between or among entities based on how it is described in text. Typically this form includes the categories of the involved entities and a label denoting the nature of their relationship, such as “symptom sign of disease” or “test reveals problem”. For example, the phrase “progressive dyspnea and lower extremity edema concerning for new-onset acute decompensated heart failure” from the last paragraph in Figure 1 contains two different “symptom sign of disease” relations. Relation extraction is usually framed as a text classification problem in which sentences or dependency paths (see sidebar) are classified into groups corresponding to relational labels. It is related to the task of knowledge base creation, which represents text as a network of structured relations over which inference can be performed to generate new knowledge (124).

Although ordinarily discussed alongside other information extraction tasks, such as NER, relation extraction is arguably one step closer to true language understanding. NER and text classification simply label text; they do not address compositionality, the combining of individual facts to generate composite ideas. Compositionality presents a particularly important challenge for clinical text mining because clinical writing reflects a high level of assumed knowledge, as well as unstated implications about the temporal and causal ordering of events. Current clinical text mining systems possess no ability to reason, as a
human would, about the relationships between laboratory and clinical findings and specific diagnoses or treatments (in Figure 1, the meaning of a clear chest x-ray or the implication of pitting edema for a diagnosis of heart failure). Such reasoning will require incorporation of external knowledge derived from, e.g., textbooks or research articles. Relation extraction is a first step in this direction.

7.1. Methods for Clinical Relation Extraction

Modern clinical relation extraction systems are generally based on deep learning models, such as CNNs with pre-trained word2vec embeddings (125), segment CNNs (Seg-CNNs) (126), and coupled Bi-LSTMs with CNNs incorporating dependency path features (127), or other machine learning methods like SVMs (128, 129). They are typically built and evaluated using annotated corpora, such as the relation extraction corpus from the 2010 i2b2/VA dataset (46), which we have seen earlier; indeed, the five studies just mentioned all used this dataset. The recent 2018 n2c2 shared task on adverse drug event relations (130) provides a recent snapshot of the field; of the top 10 systems, five used deep learning, three used SVMs, one used a random forest and one used a rule-based algorithm.

One particular relational class that has been the focus of considerable research in recent years are temporal relations, reviewed in detail in (131). A standard language has been developed for annotating temporal relations in text, including events (EVENTs), time expressions (TIMEXs), and relations between EVENTs and TIMEXs (TLINKs). This formalism has led to the creation of two major annotated corpora for clinical temporal relation extraction: the THYME corpus (132), and the 2012 i2b2 temporal relations corpus (110). Methods for temporal relation extraction have followed those developed for other clinical relation extraction tasks; earlier papers used models such as CRFs and SVMs (133), while later papers apply deep learning approaches such as CNNs (134), Bi-LSTMs (135), and BERT (136).

7.2. Inference and Entailment

Natural language inference (NLI) is a variant of relation extraction with a longstanding presence in NLP, the goal of which is to determine whether one statement (the hypothesis) can be inferred from another (the premise). As of 2018, the clinical NLP community lacked any annotated corpora for NLI, owing in part to the difficulty and expense of getting medical experts to produce annotations and the inability to share patient data with non-expert (e.g. crowd-worker) annotators. However, Romanov and Shivade (137) recently produced the MedNLI dataset to facilitate NLI research in the clinical domain. Starting with premises from the MIMIC-III (87) dataset, physicians were asked to write sentences that (1) were definitely implied by the premise, (2) were neither contradicted nor implied by the premise, and (3) were definitely contradicted by the premise. Although the task is still in its infancy, shared tasks built around the MedNLI dataset have led to multiple new approaches for NLI in this domain, including BERT-BiLSTM-Attention architectures (138), and state-of-the-art ESIM (Enhanced Sequential Inference Model) architectures coupled with knowledge-enhanced word representations based on UMLS (139, 140).
A REVIEW OF REVIEWS

The field of clinical text mining has been extensively reviewed in prior articles. The reviews selected below are those I found to be particularly useful surveys of specific research areas or the field in general.

<table>
<thead>
<tr>
<th>Year</th>
<th>Author(s)</th>
<th>Title</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>Chapman et al</td>
<td>Overcoming Barriers to NLP for Clinical Text: The Role of Shared Tasks and the Need for Additional Creative Solutions</td>
<td>(141)</td>
</tr>
<tr>
<td>2016</td>
<td>Ford et al</td>
<td>Extracting Information from the Text of Electronic Medical Records to Improve Case Detection: A Systematic Review</td>
<td>(142)</td>
</tr>
<tr>
<td>2017</td>
<td>Kreimeyer et al</td>
<td>Natural Language Processing Systems for Capturing and Standardizing Unstructured Clinical Information: A Systematic Review</td>
<td>(8)</td>
</tr>
<tr>
<td>2019</td>
<td>Khattak et al</td>
<td>A Survey of Word Embeddings for Clinical Text</td>
<td>(143)</td>
</tr>
<tr>
<td>2019</td>
<td>Mujtaba et al</td>
<td>Clinical Text Classification Research Trends: Systematic Literature Review and Open Issues</td>
<td>(20)</td>
</tr>
<tr>
<td>2020</td>
<td>Spasic et al</td>
<td>Clinical Text Data in Machine Learning: Systematic Review</td>
<td>(23)</td>
</tr>
<tr>
<td>2010</td>
<td>Stanfill et al</td>
<td>A Systematic Literature Review of Automated Clinical Coding and Classification Systems</td>
<td>(94)</td>
</tr>
<tr>
<td>2018</td>
<td>Velupillai et al</td>
<td>Using Clinical Natural Language Processing for Health Outcomes Research: Overview and Actionable Suggestions for Future Advances</td>
<td>(144)</td>
</tr>
<tr>
<td>2018</td>
<td>Wang et al</td>
<td>Clinical Information Extraction Applications: A Literature Review</td>
<td>(17)</td>
</tr>
<tr>
<td>2020</td>
<td>Wu et al</td>
<td>Deep Learning in Clinical Natural Language Processing: A Methodical Review</td>
<td>(22)</td>
</tr>
</tbody>
</table>

8. CONCLUSION

The volume of electronic health record (EHR) data in the United States has skyrocketed in recent years. In 2008, only 42% of office-based physicians reported access to an EHR; this number had climbed to 86% a decade later (145). A favorable policy environment, created by the HITECH Act of 2009 and fueled by the 21st Century Cures Act of 2016, has promoted the meaningful use of EHRs and other observational health data to inform patient care, improve health system operations, facilitate research, and provide “real world evidence” for FDA approval. Over this same time period, methodological advances in machine learning (22, 86), the creation of dedicated clinical text processing software (11, 30) and the seemingly continuous development of high-performing predictive and diagnostic algorithms (6) have fueled enthusiasm about the ability of data and data science to change the way we deliver healthcare.

Amidst such excitement, it would be easy to overlook the fact that most predictive models built on EHR data have focused on outcomes captured in structured data fields, such as mortality, readmissions, length of stay, and diagnosis codes (146, 83). In addition,
a recent systematic review found no net performance benefit of more sophisticated machine
learning methods over logistic regression in clinical prediction models (147). Both of these
observations can be explained, in part, by the limitations of clinical text mining. To date,
the vast quantities of text contained within EHRs have primarily been treated as a source of
features for downstream learning algorithms, improving predictive performance over struc-
tured data alone (148, 149, 142), but not enabling the types of fundamentally new studies
that would result if clinical text mining systems could reason about text and incorporate
prior knowledge the way a human would. Assessing whether a treatment failed or succeeded
for a given patient, for example, is still a nearly impossible task to accomplish using EHR
data without manual chart review. Even the most cutting-edge healthcare data science
companies still employ human curators to extract this type of information from text. This
situation limits both the types of questions we can ask of EHR data and the potential of
even the most sophisticated predictive algorithms and causal models to answer them.

Modern clinical text mining systems have accomplished a great deal. They can now
reliably tag a wide variety of clinically-relevant entities in text, map them to standard
concepts from lexicons and ontologies, detect negation and uncertainty, and understand
the person or people to whom they refer. Given sufficient training data, there are now
established system architectures for performing tasks like text classification and relation
extraction in the clinical domain. Production-grade clinical text mining systems are in use
throughout industry and academia, finding wide application in health outcomes research
(144), case detection and phenotyping (142), and automated coding and classification (94).
Modern deep learning methods, particularly massive language models like BERT and GPT-
3, have recently entered the clinical domain, improving state-of-the-art performance on
a variety of clinical NLP tasks and rightfully generating much excitement (86). There
remain open questions about the fundamental limitations of these methods to process and
understand language (150), and to date the rate of publications describing the application
data mining to EHR data has not kept pace with the field of EHR data mining as a whole
(17). However, the field of clinical text mining is also at an exciting turning point, as it is
beginning to pursue questions of inference and logic that cut to the heart of what it means
to build intelligent machines.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that
might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

My most sincere thanks go to Cindy Gao for reference checking and editing, Edwin Yoo
for writing the fake outpatient progress note that is included in this review, and Guergana
Savova, the creator of cTAKES, for her advice on implementing the cTAKES default
pipeline. I would also like to acknowledge Tudor Achim, a contributor to Stack Overflow,
who was the source of the clear distinction between weak and distant supervision used in
the paper (https://stackoverflow.com/questions/18944805). Thanks also to Marshall
Pierce and Ben Glicksberg, who informally reviewed the first draft of the paper and provided
helpful feedback on both the writing and the content.
LITERATURE CITED


141. Chapman WW, Nadkarni PM, Hirschman L, D’avolio LW, Savova GK, Uzuner O. 2011. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions


145. Office of the National Coordinator for Health Information Technology. 2019. ‘Office-based Physician Electronic Health Record Adoption,’ Health IT Quick-Stat #50. Tech. rep., Office of the National Coordinator for Health Information Technology


