A New Public Corpus for Clinical Section Identification: MedSecId

Paul Landes[†], Kunal Patel[‡], Sean S. Huang[◇],
Adam Webb[‡], Barbara Di Eugenio[†], and Cornelia Caragea[†]

[†]Department of Computer Science, University of Illinois at Chicago
[‡]Department of Emergency Medicine, University of Illinois at Chicago
[◇]Department of Internal Medicine and Geriatrics, University of Illinois at Chicago
{plande2,kpate318,sshuang,awebb25,bdieugen,cornelia}@uic.edu

Abstract

The process by which sections in a document are demarcated and labeled is known as section identification. Such sections are helpful to the reader when searching for information and contextualizing specific topics. The goal of this work is to segment the sections of clinical medical domain documentation. The primary contribution of this work is MedSecId, a publicly available set of 2,002 fully annotated medical notes from the MIMIC-III. We include several baselines, source code, a pretrained model and analysis of the data showing a relationship between medical concepts across sections using principal component analysis.

1 Introduction

Most unstructured medical text found in electronic health record systems (EHRs) written by medical staff have conceptually well defined sections. For example, discharge summaries are technical medical documents, written by physicians when the patient is discharged, which describe the patient's hospital stay and surrounding circumstances of their illness. As shown by the example in Figure 1, discharge summaries consist of named sections, typically in a specific sequence, such as the *History of Present Illness*; this type of section appears both in discharge summaries and in physician notes that describe a chronology of an illness that begins with the admission of the patient.

Whereas sections often have headers, section identification (SI) is more challenging than simply parsing the first several leading header tokens of the respective section (underlined in Figure 1). While the first several tokens can be helpful in identifying a section, their naming often varies. For example, the 6th section in Figure 1 starts with header tokens *Preoperative Laboratory Data*, but the section type is `labs-imaging`. There are also cases where the header tokens are missing, as shown in the 7th section (`hospital-course`) in the same figure, or where sections have several header text spans placed throughout the section. Adding to this challenge is the non-uniformity of the text, lack of section boundary syntax, and copy-pasted text from other notes or from structured data such as patient vital signs.

Admission Date: [**2126-2-7**] Discharge Date: [**2126-2-20**]
Date of Birth: [**2069-4-1**] Sex: M

history-of-present-illness
HISTORY OF PRESENT ILLNESS: Mr. [**Known last-name **] is a 56-year-old male who experienced chest...

past-medical-history
PAST MEDICAL HISTORY: Hypertension, former smoker with a 4- pack per day history for which he...

social-history
SOCIAL HISTORY: He lives alone, and he works at [**Hospital3 2576**] as a cargo transporter.

medication-history
MEDICATIONS ON ADMISSION: Aspirin 325 mg p.o. once a day, Toprol-XL 50 mg p.o. once a day.

allergies
ALLERGIES: He had no known drug allergies.

labs-imaging
PREOPERATIVE LABORATORY DATA: White count 6.0, hematocrit 33.3, platelet count 329,000...

hospital-course
On exam he had a left facial droop, status post his childhood polio. Temperature of 97.5, heart rate 65 in sinus...

discharge-diagnosis
DISCHARGE DIAGNOSES:
1. Status post coronary artery bypass grafting x 3...

discharge-instructions
DISCHARGE INSTRUCTIONS: He was instructed to make an appointment...

discharge-medications
MEDICATIONS ON DISCHARGE:
1. Aspirin enteric coated 81 mg p.o. once a day.
2. Colace 100 mg p.o. twice a day...

He was discharged to home with VNA services in good condition on [**2126-2-20**]...

Figure 1: A MIMIC-III discharge summary note with section type in **bold**, header tokens underlined, and text not belonging to any section grayed out; omitted text is indicated with ellipses.

Proceedings of the 29th International Conference on Computational Linguistics, pages 3709–3721
October 12–17, 2022.

While discharge summary sectioning helps a physician locate specific information, the primary impetus for the structure and content stems from the ongoing dispute between providers and healthcare insurance companies in the United States. Providers are limited by how much they can bill for relatively simple medical procedures, but increasingly complex procedures garner more revenue with proper documentation. Specifically, medical billing staff and insurance companies use relative value units (RVUs), which is a monetary unit updated annually and currently set at $34.30. The number of RVUs billed is based on the composition and number of sections included in the medical notes per guidelines set by the Centers for Medicare and Medicaid Services[1].

For this reason, providers are encouraged to write medical notes to maximize RVUs out of necessity (Barnes et al., 2008) even though physician training lacks such emphasis. In contrast, medical residents are evaluated with the objective structured clinical examination (OSCE), which is a student examination that evaluates students based on direct observation (Zayyan, 2011). However, the exam's evaluation with respect to medical note authoring and structure uses a very different criteria and omits RVUs (Gallagher et al., 2020). The necessity of a particular structure in medical notes, for the purpose of patient care and arguably more important insurance billing requirements, highlights the need for understanding sectioning.

However, the motivation for understanding SI is not limited to the medical field, it has a bearing on other medical NLP tasks. Since each section contains specific information, SI is often the first step in a medical NLP pipeline and can lead to downstream propagation errors causing poor task specific results if not properly executed. Examples of downstream tasks that benefit from SI include medical summarization, entity linking and natural language understanding and extraction.

While academic text segmentation has garnered interest (Hirohata et al., 2008), no publicly available medical SI annotated corpora exists (Pomares-Quimbaya et al., 2019). For this reason, we believe MedSecId is the first medical section identification dataset. It was created from 2,002 medical notes annotated by two attending physicians and one senior resident physician at the University of Illinois Chicago (UI Health). The annotation dataset is comprehensive with 2,558K annotated tokens or 97.3% of the entire corpus (see Table 1).

Description	Count
Documents	2,002
Annotations	22,561
Annotated Sentences	259,286
Total Tokens	2,630,525
Annotated Tokens	2,558,219

Table 1: Annotation dataset statistics.

The contributions of this work include: a) a comprehensive publicly available medical section annotation dataset, b) baselines with three models and several contextual and non-contextual word embeddings, c) an ontology of note to section relationships, d) human readable descriptions of medical notes and all sections annotated (see Appendix A), e) a pretrained model for each baseline, f) code to reproduce the results and read the annotations, and g) a command line tool to predict note annotations using any of the baseline models.

2 Related Work

Sectioning MedLINE abstracts was explored by McKnight and Srinivasan (2003) using a support vector machine (SVM). This classifier was used to label sentences as *Introduction*, *Method*, *Result*, or *Conclusion* and showed promising results using a bag-of-words approach. Sequence based approaches (Hirohata et al., 2008) were also used to section scientific abstracts into *Objective*, *Methods*, *Results*, and *Conclusion* labels using a conditional random field (CRF) model producing a sentence level accuracy of 95.5%.

While academic abstract segmentation was a well explored area (Hirohata et al., 2008), Tepper et al. (2012) were the first to apply statistical methods to the medical domain to automatically classify sections of clinical free text into sections. Their method used in, out, begin (IOB) annotation (Ramshaw and Marcus, 1995) with labels to mark named sections. For example, `B-HPI` indicates a beginning token for the *History of Present Illness* section. Their dataset consisted of annotating the 2010 i2b2 corpus with a section header and medical ontology label, and obtained an F-measure of 0.92 for the concept extraction task (Uzuner et al., 2011; de Bruijn et al., 2010). A Maximum Entropy (MaxEnt) model (Berger et al., 1996) and beam search were used for classification to produce the IOB sequence for token tagging.

[1] https://www.cms.gov/Regulations-and-Guidance

Along with MaxEnt, other non-neural network methods, such as SVM and CRF models continue to be popular with few exceptions as detailed in the comprehensive survey of Pomares-Quimbaya et al. (2019). One such exception (Sadoughi et al., 2018) used a long-short term memory (LSTM) model with word-to-vector (word2vec) embeddings (Mikolov et al., 2013a,b) for a binary classification of section boundaries. Even though the corpus consists of dictated and transcribed notes, they show that neural methods work for the section segmentation task. Other notable neural network (NN) text segmentation works use convolutional neural networks (CNNs) over sentence embeddings with a softmax over the output of a bi-directional long-short term memory (BiLSTM) layer to demarcate sections as a binary classification across both medical and non-medical datasets (Badjatiya et al., 2018). Barrow et al. (2020) also used a LSTM in a network that aggregates features across fastText word embeddings using a concatenated segment pooling LSTM (S-LSTM) for non-medical Wikipeda articles (Bojanowski et al., 2017).

The work of Nair et al. (2022) most closely resembles our SI work. However, their model classifies only the four SOAP (Subjective, Objective, Assessment and Plan) sections available in the corpus leaving the others as future work. Their methods also only have been tested against the Flair framework, which uses concatenated static word embeddings that are fine tuned locally for the task on the 2010 i2b2 corpus. Our method includes fine-tuning the BERT embeddings themselves as an end-to-end joint learning process. Additionally, they have provided no annotation pipeline or process to create a semi-supervised or bootstrapped corpus. Our work includes medical domain specific experiments with various word embedding combinations and novel data analysis using the Unified Medical Language System (UMLS) (Bodenreider, 2004) and *cui2vec* (Beam et al., 2020) (see Section 3.3). It also includes other methods and network experimental configurations the authors have not yet tried as they used the Flair framework "out of the box". Another significant difference is their annotations are not available[2] while we classify 50 sections and provide our code with annotations publicly.

[2]The authors did not respond to our request for obtaining their corpus for baseline comparison.

3 Dataset

MedSecId is a subset of the MIMIC-III version 1.4 corpus (Johnson et al., 2016) that we annotated; MIMIC-III is publicly available[3] and consists of critical care unit EHR records from the Beth Israel Deaconess Medical Center in Boston, Massachusetts. The dataset contains 58,976 hospital admissions across 46,520 patients who were admitted to the intensive care unit (ICU) surgical, medical, and neonatal departments. It includes 2,083,180 unstructured medical text notes handwritten by medical professionals across several disciplines and contains 15 categories, such as discharge summaries and radiology notes.

We created a curated annotation set consisting of text spans taken from a random sample across five categories of MIMIC-III medical notes[4], including discharge summaries, *Radiology*, *Consult* , *Echo*, and *Physician* progress notes (see Table 2). Each text span contains the type of the section, such as *History of Present Illness*, with zero-index character offsets of where the span starts and ends in the note.

Category	Count	Proportion
Discharge summary	1,254	62.64%
Physician	288	14.39%
Radiology	205	10.24%
Echo	198	9.89%
Consult	57	2.85%
Total	2,002	100%

Table 2: Annotated medical notes by category and their distribution in the annotation set.

While each section contains a single type, sections have zero or more overlapping header text spans (see Figure 1). In most cases, there is a single header span, but vital signs sections can "float" without a physical exam header. These header spans consist of text that identify the section such as *History Of Present Illness*, an alternate spelling or abbreviation such as *HPI*. Even though single header spans usually appear at the beginning of a section, additional section headers are found later in the body indicating subsections in some cases. Since section type inclusion highly varies based on the patient's age, notes were annotated with an age type (adult, pediatric or neonatal patient), based on the content of the note by our annotator.

[3]Access to the MIMIC-III corpus requires creating a PhysioNet account and finishing a training course.

[4]The unstructured medical note data was taken from the NOTEEVENTS table.

Type	Tokens	Spans	Notes
physical-examination	203K (8%)	1,385 (6%)	Consult, Physician
history-of-present-illness	239K (9%)	1,348 (6%)	Consult, Discharge summary, Physician
allergies	9,221 (0%)	1,205 (5%)	Consult, Discharge summary, Physician
hospital-course	692K (26%)	1,165 (5%)	Discharge summary
labs-imaging	416K (16%)	1,155 (5%)	Consult, Discharge summary, Physician
past-medical-history	60K (2%)	1,141 (5%)	Consult, Discharge summary, Physician
discharge-condition	14K (1%)	1,132 (5%)	Discharge summary
discharge-instructions	183K (7%)	1,077 (5%)	Discharge summary
discharge-diagnosis	34K (1%)	1,040 (5%)	Discharge summary
chief-complaint	9,622 (0%)	996 (4%)	Consult, Discharge summary, Physician
discharge-medications	196K (7%)	914 (4%)	Discharge summary
social-history	28K (1%)	912 (4%)	Consult, Discharge summary, Physician
medication-history	49K (2%)	867 (4%)	Consult, Discharge summary, Physician
family-history	11K (0%)	802 (4%)	Consult, Discharge summary, Physician
discharge-disposition	5,602 (0%)	754 (3%)	Discharge summary
major-surgical-or-invasive-procedure	16K (1%)	704 (3%)	Discharge summary
facility	2,668 (0%)	502 (2%)	Discharge summary
reason	5,588 (0%)	458 (2%)	Consult, Radiology
findings	58K (2%)	395 (2%)	Echo, Radiology
assessment-and-plan	131K (5%)	381 (2%)	Consult, Physician
review-of-systems	7,422 (0%)	329 (1%)	Consult, Discharge summary, Physician
image-type	1,820 (0%)	328 (1%)	Radiology
last-dose-of-antibiotics	3,689 (0%)	293 (1%)	Consult, Physician
24-hour-events	16K (1%)	250 (1%)	Physician
code-status	1,879 (0%)	237 (1%)	Physician
impression	8,233 (0%)	224 (1%)	Echo, Radiology
disposition	1,161 (0%)	210 (1%)	Physician
conclusions	28K (1%)	206 (1%)	Echo
communication	1,304 (0%)	199 (1%)	Physician
patient-test-information	13K (1%)	198 (1%)	Echo

Table 3: The top 30 most frequently annotated sections.

3.1 Annotation Process

Our annotation process consisted of several preliminary rounds of annotation, that led to our final annotation guidelines and final annotation.

Before annotation began, a custom set of regular expressions were used to pre-annotate, similar to previous work (Shivade et al., 2015); ours were medical note specific and captured header tokens along with the section spans. The application of the regular expressions was only a means to reduce the work of the annotators, who followed the annotation guide regardless of any rule based pre-annotations. The initial rule based automatic annotation process was amended by the work of Alsentzer and Kim (2018), who generously shared their *History of Present Illness* annotations to better identify and segment the initial dataset used by our annotators. These automatic annotations were edited by the annotators after they were imported into INCEpTION (Klie et al., 2018) and saved to later compute an inter-coder agreement between the physicians and rule-based output (see Section 3.2).

An attending physician (designated as a primary annotator) co-wrote a preliminary annotation guide with input from a secondary physician annotator. These two annotators engaged in a process of annotation, discussion and revision of the guidelines: they annotated a first set of one hundred notes, revised the guidelines, annotated a second set of one hundred notes, and finalized the guidelines after this second round.

Here we summarize the issues that the annotators faced during these preliminary rounds of annotation. This process was useful for the physicians to reach a consensus on what sections should be annotated and agreed on section types given their experience writing such notes themselves. A set of sections and their relation to notes began to coalesce during this process, which provided the motivation to create an ontology for the purpose of a meta documentation about the annotations and the utilitarian purpose to assist in annotation by importing it as a "knowledge base" in INCEpTION. The ontology consisted of a one-to-many mapping from notes to 50 section types using each section's header tokens captured by the regular expressions by string massaging. For example, *History of Present Illness* became `history-of-present-illness`. Among the categories, 29 sections were shared

across more than one note, such as *History of Present Illness* shared between notes *Discharge summary*, *Consult*, and *Physician* (see Table 3 for annotated sections and Appendix A for full listings).

	A1	A2	A3	R
A1	1.0	0.81	0.87	0.73
A2		1.0	0.84	0.49
A3			1.0	0.53
R				1.0

Table 4: Krippendorff's α coefficient of interannotator agreement between the annotators and the regular expressions. **A1** is the primary annotator, **A2** is the secondary annotator and **A3** is the third annotator, and **R** represents the regular expressions.

Each section type was then agreed on by the physicians with many re-typed and regrouped. For example, *Echo* notes contained internal subsections for each chamber of the heart, and was resolved by grouping the entire section as *Findings* to match section types in *Radiology* notes. Other subsections implicitly resulted by physicians copying radiology findings in discharge summaries. In an effort to reduce complexity, a flat note-to-section hierarchy without creating a second section level was kept. In some cases this was achieved by combining laboratory results data with radiology findings/diagnosis as a single section by simply re-casting *Labs* to *Labs/Radiology* for sections that included imaging studies. Other sections needed to be combined as not all notes had a clean separation.

To accommodate for a significant variation in how physicians labeled sections in these situations, *Labs* and *Radiology* was combined into a *Labs/Radiology* section. *Labs* and *Imaging* were also combined into *Labs/Imaging*. Since discharge summaries typically incorporate instructions for the patient and follow up information, we categorized these together broadly as *Discharge instructions*. The MIMIC-III pseudo tokens, such as `[**First Name**]` were not annotated unless they were included in the body of the section.

The primary and secondary annotators finished revising the annotation guidelines and then trained the third annotator. A subset of 80 medical notes, chosen from the second batch of 100 that the primary and secondary annotator had annotated and discussed, was used to train the third annotator. Because these first two batches were only used for creating guidelines and training, they were not added in the final annotation set. During this process, the well known Krippendorff's α coefficient (Krippendorff, 2011), was used to compute inter-annotator agreement (IAA) between this last annotator and the other two, until α became higher than 0.8.

3.2 Final Annotation and IAA Computation

Once the guidelines were finalized the final annotation process started. A set of 100 notes (different than the sets discussed in Section 3.1) was held out to compute the inter-annotator agreement (IAA) on the final guidelines. The remaining 1,902 notes were divvied up among the three annotators, as customary.

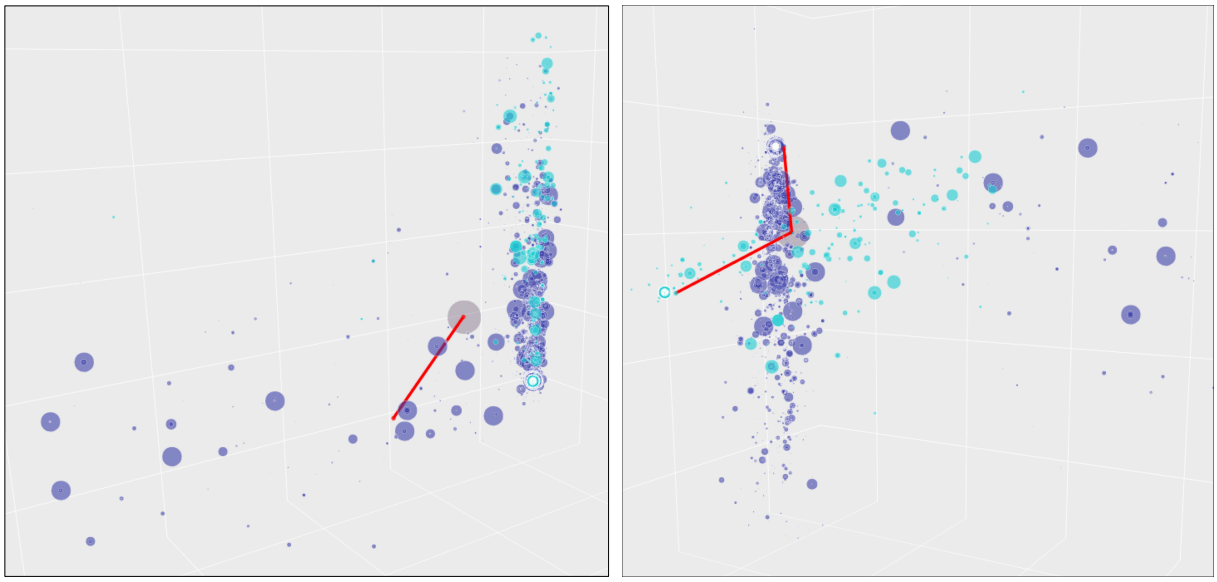
Inter-annotator agreement was calculated on the 100 held out notes as exact section character offsets and section types—both the offsets and the section type had to match to be considered correct. This agreement was calculated among the human annotators, and subsequently between each annotator and the regular expressions that were initially used to segment the notes.

Among humans, Krippendorff's α yielded more than acceptable values of 0.84 to 0.87 on the final set held out for this IAA calculation (see Table 4). At this point, these annotations were added to the final dataset by selecting notes with the fewest issues[5] using the primary annotator as the tie-breaker.

While we achieved a high inter-coder agreement among human annotators, we found troubling data in terms of the performance of the regular expression annotation approach. We computed an aggregate Krippendorff's α=0.54 between the human physician annotators and our custom regular expressions (see Section 3.1) on the final annotated data, which falls more than 14 points shy of the "lowest conceivable limit" of 0.68 (Krippendorff, 2004). This shows how regular expression's performance to segment notes falls short of that by human annotators (see Table 4), yet regular expressions continue to be the most common methods used for section identification (Pomares-Quimbaya et al., 2019; Shivade et al., 2015).

In part, the regular expressions often failed to demarcate the entire section, especially in text with irregular formatting toward the end. Furthermore, additional analysis shows the α scores between individual annotators and the regular expressions are low as well, albeit with a fairly high variance. Krippendorff suggests that acceptable scores that are

[5]Issues included placement of header tokens and missing sections. For example, an annotation with a defined section would win over another's annotation with the section type.



a) b)

Figure 2: Concept unique identifier (CUI) Plots: a) plot of `past-medical-history` (purple) and `past-surgical-history` (blue) reduced to 3D together as one data set with the first principal component (red line) with data point size as the TF/IDF score, b) plot of the same sections but reduced to 3D as separate data points with respective first principal components.

"customary to require" have $\alpha > 0.8$ (Krippendorff, 2004). On one hand, an α of 0.73 between physician *A1* and the automatic regular expression annotator *R* clears the minimal limit threshold. However, this metric falls well below the "aimed" score of 0.8. The larger issue is with physician *A2*'s and *A3*'s scores of 0.49 and 0.53, which fall short of the minimum limit by a large margin. From these scores (see Table 4) and the low overall α, we conclude regular expressions do not sufficiently segment medical notes, therefore the annotation set we provide should be considered the gold standard for medical note identification and segmentation.

3.3 Data Analysis

An interesting discovery concerned projections of medical conditions across sections in embedded space. Concept unique identifiers (CUIs) were extracted using MedCAT (Kraljevic et al., 2021) and weighted by TF/IDF (Sparck Jones, 1972) across sections. Each CUI was mapped to a vector from *cui2vec* embeddings, and then reduced to three dimensions using principal component analysis (PCA), shown in Figure 2. The plot was generated without normalizing or standardizing the data so CUI vector magnitudes were retained for analysis. Figure 2 (a) shows the `past-medical-history` section (purple) CUIs on the horizontal axis with `past-surgical-history` (blue) CUIs only on the vertical axis with size proportional to TF/IDF.

The past surgical and medical history sections in discharge summary notes project many medical disease CUIs as orthogonal to surgical CUIs. The medical disease CUIs on the vertical axis are those that do not have surgery as a treatment option, such as hypertension. However, a CUI representing coronary artery disease that plots along the surgical history vertical axis does require surgery. Most of the data points that share the vertical axis along with `past-medical-history` are those that require both medication and surgery, such as cancer.

Not only does this show *cui2vec* being used in practice for the first time, it illustrates an application of how groupings of concepts can be visualized and analyzed to gain intuition and insight in complex medical data. In our data, this includes not only a semantic relationship between concepts, but how those concepts represent the treatments involved based on the section from which they originate. Given this data relationship, we hypothesize that utilization of *cui2vec* embeddings, such as concatenating them to word vectors, will increase performance of task specific models including SI.

3.4 Limitations

MedSecId is limited to notes (with the exception of the discharge summary) of patients admitted to an ICU from the MIMIC-III corpus for several

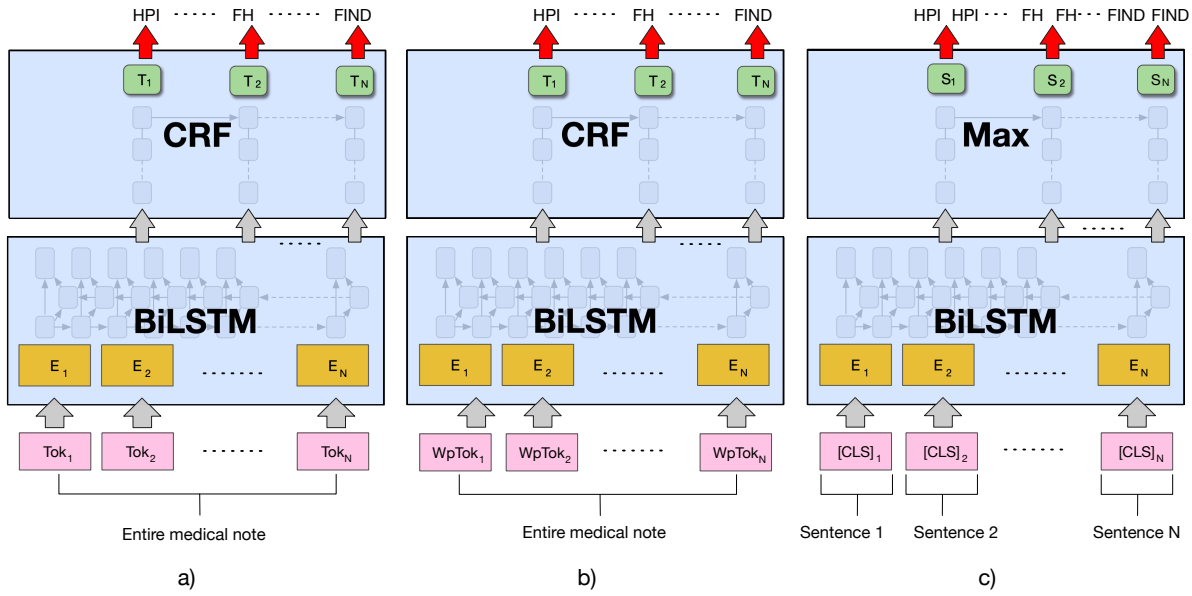


Figure 3: Baseline models: a) BiLSTM-CRF$_{tok}$ BiLSTM model with non-contextual token input embeddings, b) BERT-CRF$_{tok}$ BiLSTM model with BERT word piece token fixed input embeddings, c) BERT$_{sent}$ BiLSTM model with `[CLS]` sentence embeddings using the per sentence majority label.

note categories with no data included after the patient leaves to a lower severity department[6]. Notes written afterward are an essential source of data that provides an aspect of the patients' stay that is otherwise lacking in the corpus, such as daily progress *Physician* notes. However, *Radiology* and *Echo* notes from the MIMIC-III corpus apply to all hospital departments since they are uniform for all patients, regardless of their location, outpatient or inpatient. In addition, discharge summaries entail the entire hospital visit, including the ICU and the remainder of the admission.

3.5 Implementation Details

The annotation set was randomly sampled per note and divided as a stratified dataset into training (80%), validation (10%) and test (10%) datasets. The medical note structure ontology (see Section 3.1) is distributed as both a RDF Turtle file and a CSV file along with the annotations. The publicly available[7] code to train, validate, and test the model also includes additional APIs to access the annotated data, perform inference with the pretrained model or train a new model. This codebase includes functionality to use the pretrained model or utilize the annotations for experimentation and is ready to easily be installed.[8] This codebase also references a related project useful for parsing MIMIC-III text, pseudo token replacement, and Postgres database to Python object relational mapping.

4 Methods

Because the section text spans do not break on tokens, we cast our task as a named entity recognition (NER) using in, out (IO) encoding[9] on a 50 way classification including `<none>` for text with no sections (see Table 7). Using this encoding, we created several baselines across two BiLSTM models[10] for the purpose of future work benchmarking. These baselines include majority label metrics, a token BiLSTM-CRF, and a sentinel BERT embedding (Devlin et al., 2019) LSTM model (see Figure 3). Aside from adjusting the LSTM hidden size, gradient clipping, and number of epochs, all parameters were held constant across all experiments (see Appendix B for all hyperparameters used).

BiLSTM-CRF$_{tok}$ The token model consists of a simple non-contextual input word embeddings, a LSTM layer and fully connected linear layer using a CRF output with labels assigned by the Viterbi

[6]Only five note categories are available (see Table 2).
[7]https://github.com/uic-nlp-lab/medsecid
[8]All that is required in a `pip install`. See the GitHub repo for details.
[9]IOB encoding was not used as there are no transitions from one section to another and to reduce the label count.
[10]No models use a BERT transformer, only BERT token and sentinel ([CLS]) embeddings.

Id	Name	mF1	mP	mR	MF1	MP	MR
1	Majority Label	0.023	0.023	0.023	0.0	0	0.005
2	BERT$_{sent}$	0.925	0.925	0.925	0.589	0.616	0.6
3	BiLSTM-CRF$_{tok}$ (word2vec)	0.927	0.927	0.927	0.778	0.78	0.801
4	BERT-CRF$_{sent}$	0.929	0.929	0.929	0.689	0.734	0.7
5	BERT$_{sent}$ BioBERT	0.94	0.94	0.94	0.687	0.73	0.679
6	BERT-CRF$_{sent}$ BioBERT	0.94	0.94	0.94	0.705	0.757	0.704
7	BiLSTM-CRF$_{tok}$ (GloVE 50D)	0.954	0.954	0.954	0.76	0.783	0.765
8	BiLSTM-CRF$_{tok}$ fastText	0.954	0.954	0.954	**0.796**	0.806	0.806
9	BiLSTM-CRF$_{tok}$ (GloVE 300D)	**0.955**	0.955	0.955	0.787	0.801	0.788

Table 5: Summarization of performance metrics where mF1 is the micro F1, mP is the micro precision, mR is the micro recall, MF1 is the macro F1, MP is the macro precision, MR is the macro recall.

algorithm. Several embeddings were used with this model, including word2vec (Mikolov et al., 2013a,b), Global Vectors for Word Representation (GLoVe) (Pennington et al., 2014) and fastText (Bojanowski et al., 2017) (Crawl) embeddings.

BERT$_{\text{sent}}$ To address the issue of exploding gradients, we created a sentence-based model using static BERT sentinel embeddings to lower the input length to the LSTM layer. The model assumes sections rarely break mid-sentence since every sentence is assigned one section. Sentences with more than one section annotation will lower end-to-end performance. However, 97.6% of the annotation set contains sentences with a single section for all tokens of the respective sentence as shown in Table 6. The output of the final layer of the first time step was used as the input to a LSTM. The LSTM output forwarded to a dense layer with one output neuron for each label and an output max over the label.

Unique Sections	Count	Proportion
1	253025	97.59%
2	5589	2.16%
3	589	0.23%
4	72	0.03%
5	11	0.00%

Table 6: Distribution of sentences having a single section label across all tokens of the respective sentence.

Both the standard small BERT model and BioBERT embeddings (Lee et al., 2020) are included in the baseline results (see Section 5). A ClinicalBERT baseline model (Alsentzer et al., 2019) would not provide a fair baseline metric for comparison with future works since it trained on the MIMIC-III corpus so it was excluded.

BERT-CRF$_{\text{sent}}$ Like BERT$_{\text{sent}}$, but adds a CRF layer with Viterbi assigned labels.

5 Results

The baseline models described in Section 4 were each trained until the validation loss converged, then early stopped. The results are summarized in Table 5 with label specific results in Table 7. We report performance metrics by counting correct predictions when the character span boundaries match exactly and the sections type match. If either do not match, it is counted as an incorrect prediction.

From the majority label, it's clear the models perform comparatively well as shown in the summary results in Table 5. The GLoVe model has the best micro F1 of 0.96 with the fastText model having the best macro F1 of 0.8. This 16 point spread is evident from how performance drops off for the bottom 13 section types. Many of these low performers are those that were re-casted or re-grouped (see Section 3.1), and could be regrouped to an umbrella section type like *Labs/Imaging/Radiology* if such a rigorous delineation was not necessary.

The BERT$_{\text{sent}}$ does not lag far behind, but its performance using sentinel embeddings does not capture sections as well as the token level models despite long document length. Performance significantly improved and models converged faster with the use of gradient clipping to alleviate issues of LSTM exploding gradients (Bengio et al., 1994).

6 Conclusions and Future Work

We presented MedSecId, a comprehensive dataset of 2,002 medical annotations from the MIMIC-III corpus across five note types and 50 sections. The dataset contains section types, headers and patient age annotations. Our dataset shows promising baseline results from simple models such as BiLSTMs with diverse inputs, but still leaves room for improvement by more sophisticated models.

We expect performance using our models to improve pipelines that use rule based methods for

Id	Label	mF1	mP	mR	MF1	MP	MR	Acc	Count
1	procedure	0	0	0	0	0	0	0	156
2	labs	0	0	0	0	0	0	0	436
3	prenatal-screens	0.276	0.276	0.276	0.216	0.5	0.138	0.276	105
4	imaging	0.357	0.357	0.357	0.263	0.5	0.178	0.357	990
5	comparison	0.414	0.414	0.414	0.195	0.333	0.138	0.414	222
6	code-status	0.513	0.513	0.513	0.226	0.333	0.171	0.513	150
7	wet-read	0.521	0.521	0.521	0.342	0.5	0.26	0.521	121
8	communication	0.556	0.556	0.556	0.179	0.25	0.139	0.556	133
9	impression	0.563	0.563	0.563	0.18	0.25	0.141	0.563	920
10	disposition	0.647	0.647	0.647	0.262	0.333	0.216	0.647	68
11	history	0.688	0.688	0.688	0.272	0.333	0.229	0.688	170
12	past-surgical-history	0.745	0.745	0.745	0.427	0.5	0.372	0.745	145
13	current-medications	0.746	0.746	0.746	0.142	0.167	0.124	0.746	1406
14	contrast	0.8	0.8	0.8	0.444	0.5	0.4	0.8	25
15	<none>	0.816	0.816	0.816	0.03	0.033	0.027	0.816	6378
16	discharge-disposition	0.83	0.83	0.83	0.151	0.167	0.138	0.83	513
17	addendum	0.833	0.833	0.833	0.151	0.167	0.139	0.833	3106
18	last-dose-of-antibiotics	0.872	0.872	0.872	0.466	0.5	0.436	0.872	397
19	indication	0.88	0.88	0.88	0.468	0.5	0.44	0.88	117
20	physical-examination	0.881	0.881	0.881	0.156	0.167	0.147	0.881	22113
21	image-type	0.884	0.884	0.884	0.313	0.333	0.295	0.884	181
22	discharge-condition	0.904	0.904	0.904	0.317	0.333	0.301	0.904	1490
23	infusions	0.909	0.909	0.909	0.476	0.5	0.455	0.909	99
24	history-of-present-illness	0.924	0.924	0.924	0.137	0.143	0.132	0.924	24950
25	discharge-medications	0.925	0.925	0.925	0.192	0.2	0.185	0.925	25088
26	flowsheet-data-vitals	0.932	0.932	0.932	0.482	0.5	0.466	0.932	2128
27	24-hour-events	0.954	0.954	0.954	0.244	0.25	0.238	0.954	1765
28	past-medical-history	0.959	0.959	0.959	0.163	0.167	0.16	0.959	5990
29	discharge-diagnosis	0.959	0.959	0.959	0.196	0.2	0.192	0.959	3578
30	family-history	0.968	0.968	0.968	0.328	0.333	0.323	0.968	1171
31	chief-complaint	0.968	0.968	0.968	0.492	0.5	0.484	0.968	1142
32	medical-condition	0.971	0.971	0.971	0.328	0.333	0.324	0.971	409
33	review-of-systems	0.977	0.977	0.977	0.494	0.5	0.488	0.977	724
34	labs-imaging	0.981	0.981	0.981	0.142	0.143	0.14	0.981	45855
35	discharge-instructions	0.986	0.986	0.986	0.166	0.167	0.164	0.986	23208
36	social-history	0.988	0.988	0.988	0.249	0.25	0.247	0.988	3114
37	allergies	0.989	0.989	0.989	0.331	0.333	0.33	0.989	891
38	assessment-and-plan	0.99	0.99	0.99	0.199	0.2	0.198	0.99	12728
39	reason	0.992	0.992	0.992	0.332	0.333	0.331	0.992	646
40	conclusions	0.994	0.994	0.994	0.498	0.5	0.497	0.994	2814
41	findings	0.998	0.998	0.998	0.333	0.333	0.333	0.998	6053
42	hospital-course	0.998	0.998	0.998	0.2	0.2	0.2	0.998	78321
43	social-and-family-history	1	1	1	1	1	1	1	52
44	technique	1	1	1	1	1	1	1	22
45	clinical-implications	1	1	1	1	1	1	1	36
46	other-medications	1	1	1	1	1	1	1	489
47	major-surgical-or-invasive-procedure	1	1	1	1	1	1	1	1903
48	facility	1	1	1	1	1	1	1	344
49	patient-test-information	1	1	1	1	1	1	1	1349
50	medication-history	1	1	1	0.333	0.333	0.333	1	6082

Table 7: By label BiLSTM-CRF$_{tok}$ performance where mF1 is the micro F1, mP is the micro precision, mR is the micro recall, MF1 is the macro F1, MP is the macro precision, MR is the macro recall, Acc is the accuracy, count is the the number of tokens encountered in the test set. The <none> label is for tokens with no section annotated.

SI as mentioned in Section 3.2. These pipelines include discharge note summarization, and other downstream tasks that would benefit from having header and non-section text removed such as training word embeddings such as ClinicalBERT.

Hyperparameter tuning with the baseline models is a next logical step for further work. Another obvious opportunity to improve performance is to concatenate *cui2vec* embeddings in the input layer as described in Section 3.3. Other future work includes comparing the results using the synthetic tokens in place of pseudo tokens, which would shed light on how models learn with more realistic data.

Acknowledgments

This work was supported by award R01 CA225446 from the National Institutes of Health. We thank Andy Boyd for his support on this work.

References

Emily Alsentzer and Anne Kim. 2018. Extractive Summarization of EHR Discharge Notes. arXiv: 1810.12085 (Only available as arXiv preprint).

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. pages 72–78.

Pinkesh Badjatiya, Litton J. Kurisinkel, Manish Gupta, and Vasudeva Varma. 2018. Attention-Based Neural Text Segmentation. In *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 180–193. Springer International Publishing.

Stephen L. Barnes, Bryce R. H. Robinson, J. Taliesin Richards, Cindy E. Zimmerman, Tim A. Pritts, Betty J. Tsuei, Karyn L. Butler, Peter C. Muskat, Kenneth Davis, and Jay A. Johannigman. 2008. The devil is in the details: Maximizing revenue for daily trauma care. *Surgery*, 144(4):670–676.

Joe Barrow, Rajiv Jain, Vlad Morariu, Varun Manjunatha, Douglas Oard, and Philip Resnik. 2020. A Joint Model for Document Segmentation and Segment Labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 313–322. Association for Computational Linguistics.

Andrew L. Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin Weber, Nathan Palmer, Xu Shi, Tianxi Cai, and Isaac S. Kohane. 2020. Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 25:295–306.

Y. Bengio, P. Simard, and P. Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Berry de Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu. 2010. NRC at i2b2: One challenge, three practical tasks, nine statistical systems, hundreds of clinical records, millions of useful features. In *Proceedings of the 2010 I2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*, page 5.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Benjamin D. Gallagher, Saman Nematollahi, Henry Park, and Salila Kurra. 2020. Comparing Students' Clinical Grades to Scores on a Standardized Patient Note-Writing Task. *Journal of General Internal Medicine*, 35(11):3243–3247.

A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4. IEEE.

Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. 2008. Identifying Sections in Scientific Abstracts using Conditional Random Fields. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A. Folarin, Angus Roberts, Rebecca Bendayan, Mark P. Richardson, Robert Stewart, Anoop D. Shah, Wai Keong Wong, Zina Ibrahim, James T. Teo, and Richard J. B. Dobson. 2021. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. *Artificial Intelligence in Medicine*, 117:102083.

Klaus Krippendorff. 2004. Reliability in Content Analysis. *Human Communication Research*, 30(3):411–433.

Klaus Krippendorff. 2011. Agreement and Information in the Reliability of Coding. *Communication Methods and Measures*, 5(2):93–112.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling

Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289. Morgan Kaufmann Publishers Inc.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Larry McKnight and Padmini Srinivasan. 2003. Categorization of Sentence Types in Medical Abstracts. *AMIA Annual Symposium Proceedings*, 2003:440–444.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. arXiv: 1301.3781 (Only available as arXiv preprint).

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Namrata Nair, Sankaran Narayanan, Pradeep Achan, and K. P. Soman. 2022. Clinical Note Section Identification Using Transfer Learning. In Xin-She Yang, Simon Sherratt, Nilanjan Dey, and Amit Joshi, editors, *Proceedings of Sixth International Congress on Information and Communication Technology*, volume 235 of *Lecture Notes in Networks and Systems*, pages 533–542. Springer Singapore.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Alexandra Pomares-Quimbaya, Markus Kreuzthaler, and Stefan Schulz. 2019. Current approaches to identify sections within clinical narratives from electronic health records: A systematic review. *BMC Medical Research Methodology*, 19(1):155.

Lance Ramshaw and Mitch Marcus. 1995. Text Chunking using Transformation-Based Learning. In *Third Workshop on Very Large Corpora*. ACL.

Najmeh Sadoughi, Greg P. Finley, Erik Edwards, Amanda Robinson, Maxim Korenevsky, Michael Brenndoerfer, Nico Axtmann, Mark Miller, and David Suendermann-Oeft. 2018. Detecting Section Boundaries in Medical Dictations: Toward Real-Time Conversion of Medical Dictations to Clinical Reports. In *Speech and Computer*, Lecture Notes in Computer Science, pages 563–573. Springer International Publishing.

Chaitanya Shivade, Pranav Malewadkar, Eric Fosler-Lussier, and Albert M. Lai. 2015. Comparison of UMLS terminologies to identify risk of heart disease using clinical notes. *Journal of Biomedical Informatics*, 58 Suppl:S103–S110.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Michael Tepper, Daniel Capurro, Fei Xia, Lucy Vanderwende, and Meliha Yetisgen-Yildiz. 2012. Statistical Section Segmentation in Free-Text Clinical Records. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2001–2008. European Language Resources Association (ELRA).

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):552–556.

Marliyya Zayyan. 2011. Objective Structured Clinical Examination: The Assessment of Choice. *Oman Medical Journal*, 26(4):219–222.

A Descriptions

Table 8: Note Categories

Name	Description
Consult	Notes generated when a specialst intervenes in a patient's care.
Discharge summary	A discharge summary describes a patient's stay at a hospital and the care they received. They can also include follow up instructions, medications and a schedule for future appointments.
Echo	An ultrasound of the heart.
Physician	Daily notes taken by the physician on their rounds as a part of a patient check up.
Radiology	Diagnosis and other notes taken by a radiologiest based on images such as xrays, MRI, CAT scans.

Table 9: Section Types

Section Type	Name	Description
24-hour-events	24 Hour Events	Description of what happened in the past 24 hours of the patients stay.
addendum	Addendum	An addition to the note.
allergies	Allergies	Patient allergies to medication and food of varying severity.
assessment-and-plan	Assessment And Plan	An overview of the problems that are occuring and the plan to address each problem.
critical-care-attending-addendum	Attending Addendum	The attending physician's addition to the note.
chief-complaint	Chief Complaint	The reason why the patient came to the hospital.
clinical-implications	Clinical Implications	Why this study is important.
code-status	Code Status	What should be done in the event of a cardiac or respiratory arrest, end of goals care.
communication	Communication	Information about who to contact and the relation to the patient.
comparison	Comparison	Comparing the new study to prior studies to determine interval changes.
conclusions	Conclusions	Interpretation of the findings in relation to the patient's condition.
contrast	Contrast	Was contrast introduced into the patient.
current-medications	Current Medications	Medications that the patient are taking at home.
discharge-condition	Discharge Condition	The stability of the patient upon discharge.
discharge-diagnosis	Discharge Diagnosis	The diagnosis of the patient after being worked up in the hospital.
discharge-disposition	Discharge Disposition	Where the patient is being discharged to.
discharge-instructions	Discharge Instructions	Post discharge instructions regarding what the patient can and cannot do.
discharge-medications	Discharge Medications	Medications that the patient will sent home with and to continue taking.
disposition	Disposition	Where the patient will go within the hospital.
family-history	Family History	Medical history of family members.
findings	Findings	Specific finidngs during the study.
flowsheet-data-vitals	Flowsheet Data/Vitals	Information pulled from flowsheets that are discretely kept within the ehr.
history	History	Patient's clinical history warranting exam.
history-of-present-illness	History Of Present Illness	A description of the events surrounding the reason why the patient came to the hospital: Symptom onset, duration, severity and associatating factors.
hospital-course	Hospital Course	A summary of what happened during the patient's time in the hospital.
image-type	Image Type	The type of study being performed.
imaging	Imaging	All image related orders placed by the physician including: CT, XRAY, ECHO, MRI, Ultrasound.
impression	Impression	Overall summerization of the study.
indication	Indication	Why the study was performed.
infusions	Infusions	Medications classified as a constant infusion.
labs	Labs	Laboratory values.
labs-imaging	Labs / Imaging	Lab and radiological results.
last-dose-of-antibiotics	Last Dose Of Antibiotics	Time of the last dose of antibiotic medications.
major-surgical-or-invasive-procedure	Major Surgical Or [...]	Any procedures or surgies that occured while the patient was at the hospital.
medical-condition	Medical Condition	History of the patient and why the patient needs the study.

Continued on the next page

Table 9: Section Types (cont)

Section Type	Name	Description
medication-history	Medication History	Medications that the patient are taking at home.
other-medications	Other Medications	Other medications the patient is receiving.
past-medical-history	Past Medical History	Medical problems a patient has.
past-surgical-history	Past Surgical History	All surgeries the patient has had in their past.
patient-test-information	Patient/Test Information	Basic and standardized information of the patient.
physical-examination	Physical Examination	Evalutating anatomic finds of a patient through palpation and auscultation.
prenatal-screens	Prenatal Screens	Screening of blood type and infections prior to delivery.
procedure	Procedure	Procedure name.
reason	Reason	Why the consulting team was brought in for the patient's care.
review-of-systems	Review of Systems	A generalized review of potential symptoms that the patient might not have addressed in the chief complaint or history of present illness.
social-history	Social History	History of occupation, recreational activities, and living situation.
social-and-family-history	Social and Family History	Combination of social and family history.
technique	Technique	How the procedure was being performed.
wet-read	Wet Read	Initial read, not the official read of the study.
addendum	addendum	An addition to the note.
facility	facility	The location the patient is going after discharge.

B Hyperparameters

The hyperparameters used to train the models described in Section 4. Those hyperparameters which differed for each model are given in Table 10. Hyperparameters shared across all models are given in Table 11. The only non-zero drop out was used in the LSTM layer.

Model	Epochs	Learning Rate	CRF
BERT-CRF$_{sent}$	40	0.003	True
BERT-CRF$_{sent}$ BioBERT	45	0.003	True
BERT$_{sent}$	35	0.003	False
BERT$_{sent}$ BioBERT	45	0.003	False
BiLSTM-CRF$_{tok}$ (GloVE 300D)	30	0.01	True
BiLSTM-CRF$_{tok}$ (GloVE 50D)	25	0.01	True
BiLSTM-CRF$_{tok}$ (word2vec)	30	0.01	True
BiLSTM-CRF$_{tok}$ fastText	40	0.01	True

Table 10: The hyperparameters of the models given in the results. Epochs is the number of epochs used to train the model, Learning Rate is the learning rate for the update step size of the loss function and CRF is whether the BiLSTM used a CRF output layer.

Name	Value	Description
Batch Size	20	The size of the mini-batches used to train the model.
Hidden Size	250	The hidden size of the LSTM.
Num Layers	2	The number of stacked layers of the LSTM.
Dropout	0.15	The dropout of the LSTM.

Table 11: The shared hyperparameters set for all models.