Extracting COVID-19 Diagnoses and Symptoms From Clinical Text: A New Annotated Corpus and Neural Event Extraction Framework

Kevin Lybarger[1] **Mari Ostendorf**[1] **Matthew Thompson**[2] **Meliha Yetisgen**[3]

[1]Electrical & Computer Engineering, University of Washington
[2]Department of Family Medicine, University of Washington
[3]Biomedical & Health Informatics, University of Washington
{lybarger, ostendor, mjt, melihay}@uw.edu

Abstract

Coronavirus disease 2019 (COVID-19) is a global pandemic. Although much has been learned about the novel coronavirus since its emergence, there are many open questions related to tracking its spread, describing symptomology, predicting the severity of infection, and forecasting healthcare utilization. Free-text clinical notes contain critical information for resolving these questions. Data-driven, automatic information extraction models are needed to use this text-encoded information in large-scale studies. This work presents a new clinical corpus, referred to as the COVID-19 Annotated Clinical Text (CACT) Corpus, which comprises 1,472 notes with detailed annotations characterizing COVID-19 diagnoses, testing, and clinical presentation. We introduce a span-based event extraction model that jointly extracts all annotated phenomena, achieving high performance in identifying COVID-19 and symptom events with associated assertion values (0.83-0.97 F1 for events and 0.73-0.79 F1 for assertions).

1 Introduction

As of June 22, 2020, there were over 8.9 million confirmed COVID-19 cases globally, resulting in 466 thousand related deaths (World Health Organization, 2020a). Surveillance efforts to track the spread of COVID-19 and estimate the true number of infections remains a challenge for policy makers, healthcare workers, and researchers, even as testing availability increases. Symptom information provides useful indicators for tracking potential COVID-19 infections and disease clusters (Rossman et al., 2020). Certain symptoms and underlying comorbidities have directed COVID-19 testing. However, the clinical presentation of COVID-19 varies significantly in severity and symptom profiles (Wu and McGoogan, 2020).

The most prevalent COVID-19 symptoms reported to date are fever, cough, fatigue, and dyspnea (Yang et al., 2020), but emerging reports identify additional symptoms, including diarrhea and neurological symptoms, such as changes in taste or smell (Vetter et al., 2020; Qian et al., 2020; Wei et al., 2020). Certain initial symptoms may be associated with higher risk of complications; in one study, dyspnea was associated with a two-fold increased risk of Acute Respiratory Distress Syndrome (Wu et al., 2020). However, correlations between symptoms, positive tests, and rapid clinical deterioration are not well understood in ambulatory care and emergency department settings.

Routinely collected information in the Electronic Health Record (EHR) can provide crucial COVID-19 testing, diagnosis, and symptom data needed to address these knowledge gaps. Test results can easily be queried and analyzed at scale from structured EHR data. However, more detailed and nuanced descriptions of COVID-19 diagnoses, exposure history, symptoms, and clinical decision-making are typically only documented in the clinical narrative. To leverage this textual information in large-scale studies, the salient COVID-19 and symptom information must be automatically extracted.

This work presents a new corpus of clinical text annotated for COVID-19, referred to as the COVID-19 Annotated Clinical Text (CACT) Corpus. CACT consists of 1,472 notes from the University of Washington (UW) clinical repository with detailed event-based annotations for COVID-19 diagnosis, testing, and symptoms. Given the recent rapid emergence of the pandemic, CACT is one of the first clinical data sets with COVID-19 annotations and includes 29.9K distinct events. We also present the first information extraction results on CACT using an end-to-end neural event extraction model, establishing a strong baseline for identifying COVID-19 and symptom events.

2 Related Work

2.1 Annotated Corpora

Given the recent onset of COVID-19, there are limited COVID-19 corpora for natural language processing (NLP) experimentation. Corpora of scientific papers related to COVID-19 are available (Wang et al., 2020a; World Health Organization, 2020b), and automatic labels for biomedical entity types are available for some of these research papers (Wang et al., 2020b). However, we are unaware of corpora of clinical text with supervised COVID-19 annotations.

Multiple clinical corpora are annotated for symptoms. As examples, South et al. (2009) annotated symptoms and other medical concepts with negation (present/not present), temporality, and other attributes. Koeling et al. (2011) annotated a predefined set of symptoms related to ovarian cancer. For the i2b2/VA challenge, Uzuner et al. (2011) annotated annotated medical concepts, including symptoms, with assertion values and relations.

2.2 Relation and Event Extraction

There is a significant body of information extraction (IE) work related to coreference resolution, relation extraction, and event extraction tasks. In these tasks, spans of interest are identified, and linkages between spans are predicted. Many contemporary IE systems use end-to-end multi-layer neural models that encode an input word sequence using recurrent or transformer layers, classify spans (entities, arguments, etc.), and predict the relationship between spans (coreference, relation, role, etc.) (Zheng et al., 2017; Orr et al., 2018; Shi et al., 2019; Pang et al., 2019; Chen et al., 2019; Christopoulou et al., 2020). Of most relevance to our work is a series of developments starting with Lee et al. (2017), which introduces a span-based coreference resolution model that enumerates all spans in a word sequence, predicts entities using a feed-forward neural network (FFNN) operating on span representations, and resolves coreferences using FFNNs operating on entity span-pairs. Luan et al. (2018) adapted this framework to entity and relation extraction, with a specific focus on scientific literature. Luan et al. (2019) extended the method to take advantage both co-reference and relation links in a graph-based approach to jointly predict entity spans, co-reference and relations. By updating span representations in multi-sentence co-reference chains, the graph-based approach achieved state-of-the-art on several IE tasks representing a range of different genres. Wadden et al. (2019) expands on Luan et al. (2019)'s approach, adapting it to event extraction tasks. We build on Luan et al. (2018) and Wadden et al. (2019)'s work, augmenting the modeling framework to fit the CACT annotation scheme. In CACT, event arguments are generally close to the associated trigger, and inter-sentence events linked by co-reference are infrequent, so the graph-based extension, which adds complexity, is unlikely to benefit our extraction task.

Many recent NLP systems use pre-trained language models (LMs), such as ELMo, BERT, and XLNet, that leverage unannotated text (Peters et al., 2018; Devlin et al., 2019; Yang et al., 2019). A variety of strategies for incorporating the LM output are used in IE systems, including using the contextualized word embedding sequence: as the input to a Conditional Random Field entity extraction layer (Huang et al., 2019), as the basis for building span representations (Luan et al., 2019; Wadden et al., 2019), or by adding an entity-aware attention mechanism and pooled output states to a fully transformer-based model (Wang et al., 2019). There are many domain-specific LM variants. Here, we use Alsentzer et al. (2019)'s *Bio+Clinical BERT*, which is trained on PubMed papers and MIMIC-III (Johnson et al., 2016) clinical notes.

3 Materials

3.1 Data

This work used inpatient and outpatient clinical notes from the UW clinical repository. COVID-19-related notes were identified by searching for variations of the terms *coronavirus*, *covid*, *sars-cov*, and *sars-2* in notes authored between February 20-March 31, 2020, resulting in a pool of 92K notes. This work utilized a subset of 53K notes, including only notes with at least five sentences and corresponding to one of six types: telephone encounters, outpatient progress, emergency department, inpatient nursing, intensive care unit, and general inpatient medicine. Multiple note types were used to improve the extraction model generalizability.

Early in the outbreak, the UW EHR did not include COVID-19 specific structured data; however, structured fields indicating COVID-19 test types and results were added as testing expanded. We used these structured fields to assign a *COVID-19 Test* label to each note based on the patient test status at the time of note creation:

- *none*: patient not tested
- *positive*: patient tested positive
- *negative*: patient tested negative

More nuanced descriptions of COVID-19 testing (e.g. conditional or unordered tests) or diagnoses (e.g. possible infection or exposure) are not available as structured data. For the 53K note subset, the *COVID-19 Test* label distribution is 90.8% *none*, 1.3% *positive*, and 7.9% *negative*.

Given the sparsity of *positive* and *negative* notes, CACT is intentionally biased to increase the prevalence of these labels. To ensure adequate *positive* training samples, the CACT training partition includes 50% *positive* notes and 50% *none* and *negative* notes. Ideally, the test set would be representative of the true distribution; however, the expected number of *positive* labels with random selection is insufficient to evaluate extraction performance. Consequently, the CACT test partition includes 50% *positive* and *negative* notes and 50% *none* notes. Notes were randomly selected in equal proportions from the six note types.

3.2 Annotation Scheme

We created detailed annotation guidelines for two event types, *COVID* and *Symptom*, which are summarized in Table 1. *COVID* and *Symptom* are annotated as events, where each event includes a trigger and arguments characterizing the event. *COVID* trigger is generally an explicit COVID-19 reference, like “COVID-19” or “coronavirus.” *COVID Test Status* characterizes implicit and explicit references to testing, and *Assertion* captures diagnoses and hypothetical references to COVID-19. *Symptom* events capture subjective, often patient reported, indications of disorders and diseases (e.g “cough”). *Symptom* trigger identifies the specific symptom, for example “wheezing” or “fever,” which are characterized through *Assertion*, *Change*, *Severity*, *Anatomy*, *Characteristics*, *Duration*, and *Frequency* arguments. *Labeled arguments* (e.g. *Assertion*) include an argument span, type, and subtype (e.g. *present*). *Span-only arguments*, like *Characteristics*, include an argument span and type, without a subtype label. Notes were annotated using the BRAT annotation tool (Stenetorp et al., 2012). Figure 1 presents BRAT annotation examples.

3.3 Annotation Scoring and Evaluation

Annotation and extraction is scored as a slot filling task, focusing on information most relevant to secondary use applications. Figure 2 presents the same sentence annotated by two annotators, along with the populated slots for the *Symptom* event. Both annotations include the same trigger and *Frequency* spans (“cough” and “intermittent”, respectively). The *Assertion* spans differ (“presenting with” vs. “presenting”), but the assigned subtypes (*present*) are the same, so the annotations are equivalent for purposes of populating a database. Annotator agreement and extraction performance are assessed using scoring criteria that reflects this slot filling interpretation of the labeling task.

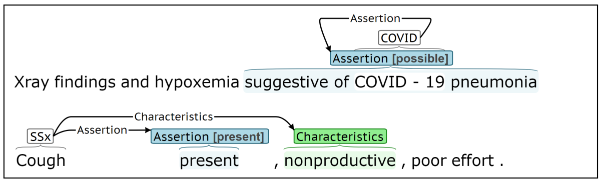

Figure 1: BRAT annotation examples for *COVID* and *Symptom* (*SSx*) event types

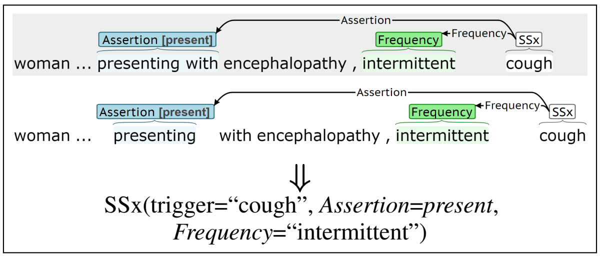
Assertion
Assertion [present]
Frequency
Frequency
SSx
woman ... presenting with encephalopathy , intermittent
cough
Assertion
Assertion [present]
Frequency
Frequency
SSx
woman ...
presenting
with encephalopathy , intermittent
cough
⇓
SSx(trigger=“cough”, Assertion=present,
Frequency=“intermittent”)

Figure 2: Annotation examples describing event extraction as a slot filling task

The *Symptom* trigger span identifies the specific symptom. For *COVID*, the trigger anchors the event, although the span text is not salient to downstream applications. For labeled arguments, the subtype label captures the most salient information, and the identified span is less informative. For span-only arguments, the spans are not easily mapped to a fixed label set, so the selected span contains the salient information. Performance is evaluated using precision, recall, and F1.

Trigger: Triggers, T_i, are represented by a pair (event type, e_i; token indices, x_i). Trigger equivalence is defined as

$$T_i \equiv T_j \text{ if } (e_i \equiv e_j) \wedge (x_i \equiv x_j).$$

Arguments: Events are aligned based on trigger equivalence. The arguments of events with equivalent triggers are compared using different criteria for *labeled arguments* and *span-only arguments*. Labeled arguments, L_i, are represented as

Event type, e	**Argument type, a**	**Argument subtypes, L_l**	**Span examples**
COVID	Trigger*	–	“COVID,” “COVID-19”
	Test Status†	{positive, negative, pending, conditional, not ordered, not patient, indeterminate }	“tested positive”
	Assertion†	{present, absent, possible, hypothetical, not patient}	“positive,” “low suspicion”
Symptom	Trigger*	–	“cough,” “shortness of breath”
	Assertion*	{present, absent, possible, conditional, hypothetical, not patient}	“admits,” “denies”
	Change	{no change, worsened, improved, resolved}	“improved,” “continues”
	Severity	{mild, moderate, severe}	“mild,” “required ventilation”
	Anatomy	–	“chest wall,” “lower back”
	Characteristics	–	“wet productive,” “diffuse”
	Duration	–	“for two days,” “1 week”
	Frequency	–	“occasional,” “chronic”

Table 1: Annotation guideline summary. * indicates the argument is required. † indicates at least one of the arguments, *Test Status* or *Assertion*, is required

a triple (argument type, a_i; token indices, x_i; subtype, l_i). For labeled arguments, the argument type, a, and subtype, l, capture the salient information and equivalence is defined as

$$L_i \equiv L_j \text{ if } (T_i \equiv T_j) \wedge (a_i \equiv a_j) \wedge (l_i \equiv l_j).$$

Span-only arguments, S_i, are represented as a pair (argument type, a_i; token indices, x_i). Arguments with equivalent triggers and argument types, $(T_i \equiv T_j) \wedge (a_i \equiv a_j)$, are compared at the token-level (rather than the span-level) to allow partial matches. Partial match scoring is used as partial matches can still contain useful information.

3.4 Annotation Statistics

CACT includes 1,472 notes with a 70%/30% train/test split and 29.9K events annotated (5.4K *COVID* and 24.4K *Symptom*). Figure 3 contains a summary of the *COVID* annotation statistics for the train/test subsets. By design, the training and test sets include high rates of COVID-19 infection (*present* subtype for *Assertion* and *positive* subtype for *Test Status*), with higher rates in the training set. CACT includes high rates of *Assertion hypothetical* and *possible* subtypes. The *hypothetical* subtype applies to sentences like, “She is mildly concerned about the coronavirus” and “She cancelled nexplanon replacement due to COVID-19.” The *possible* subtype applies to sentences like, “risk of Covid exposure” and “Concern for respiratory illness (including COVID-19 and influenza).” *Test Status pending* is also frequent.

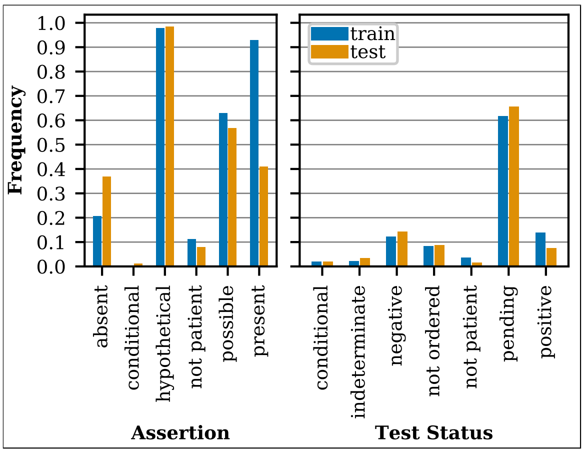

Figure 3: COVID annotation summary

There is some variability in the endpoints of the annotated *COVID* trigger spans (e.g. “COVID” vs. “COVID test”); however 98% of the *COVID* trigger spans in the training set start with the tokens “COVID,” “COVID19,” or “coronavirus.” Since the *COVID* trigger span is only used to anchor and disambiguate events, the *COVID* trigger spans were truncated to the first token of the annotated span in all experimentation and results.

The training set includes 1,756 distinct uncased *Symptom* trigger spans, 1,425 of which occur fewer than five times. The identified symptoms were not normalized to canonical forms (e.g. “shortness of breath” and “sob” considered distinct symptoms). Figure 4 presents the frequency of the 20 most common *Symptom* trigger spans in the training set by *Assertion* subtypes *present*, *absent*, and other (*possible*, *conditional*, *hypothetical*, or *not patient*). These 20 symptoms account for 49% of the training set *Symptom* events. There is ambiguity in delineating between some symptoms and other clinical phenomena (e.g. exam findings and medical problems), which introduces some annotation noise.

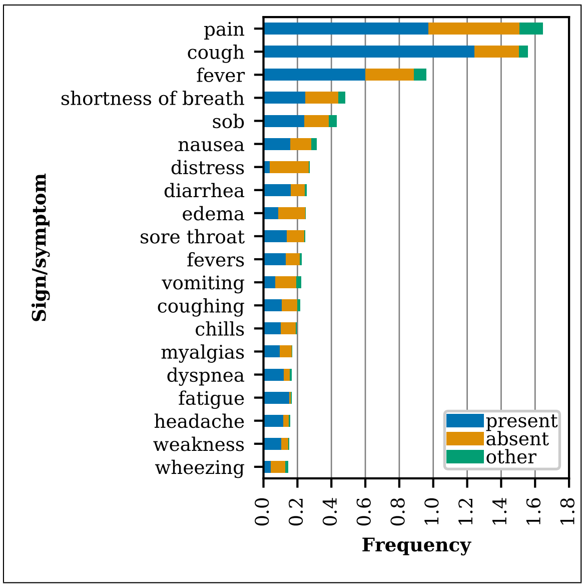

Figure 4: Most frequent symptoms in the training set broken down by *Assertion* subtype

Given the long tail of the symptom distribution and our desire to understand the more prominent COVID-19 symptoms, we focused annotator agreement assessment and extraction model training/evaluation on the symptoms that occurred at least 10 times in the training set, resulting in 185 distinct symptoms that cover 82% of the training set *Symptom* events. The set of 185 symptoms was determined only using the training set, to allow unbiased experimentation on the test set. All subsequent results and experimentation only incorporate these 185 most frequent symptoms.

3.5 Annotator Agreement

The first two rounds of annotation were doubly annotated (72 notes in *round 1* and 96 notes in *round 2*). Figure 5 presents the annotator agreement for each annotation round. For labeled arguments, F1 scores are micro-average across subtypes. After *round 1*, annotator disagreements were carefully reviewed, the annotation guidelines were updated, and annotators received additional training. Starting with *round 2*, potential *COVID* triggers were pre-annotated using pattern matching ("COVID," "COVID-19," "coronavirus," etc.), to improve the recall of *COVID* annotations. Pre-annotated *COVID* triggers were modified as needed by the annotators, including removing, shifting, and adding trigger spans. The guideline updates, additional annotator training, and pre-annotation of *COVID* triggers resulted in improved agreement across all labeled phenomena, except for *Change*.

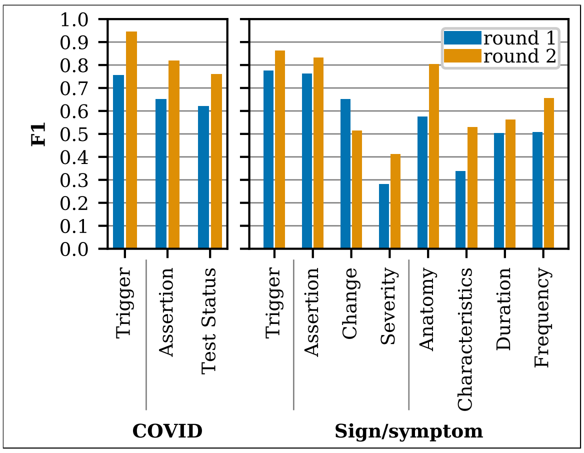

Figure 5: Annotator agreement

4 Event Extraction Model

Event extraction tasks, like ACE05 (Walker et al., 2006), typically require prediction of the following event phenomena:

- trigger span identification
- trigger type (event type) classification
- argument span identification
- argument type/role classification

The CACT annotation scheme differs from this configuration in that labeled arguments require the argument type (e.g. *Assertion*) and the subtype (e.g. *present*, *absent*, etc.) to be predicted. Resolving the argument subtypes require a classifier with additional predictive capacity.

We implement a span-based, end-to-end, multilayer event extraction model that jointly predicts all event phenomena, including the trigger span, event type, and argument spans, types, and subtypes. Figure 6 presents our extraction framework, which differs from prior related work in that multiple span classifiers are used to accommodate the argument subtypes.

Each input sentence consists of tokens, $X = \{x_1, x_2, ...x_n\}$, where n is the number of tokens. For each sentence, the set of all possible spans, $S = \{s_1, s_2, ...s_m\}$, is enumerated, where m is the number of spans with token length less than or equal to M tokens. The model generates trigger

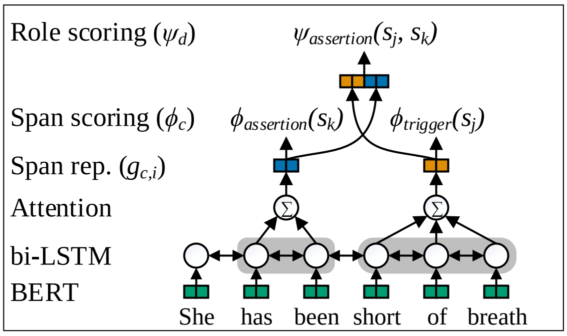

Figure 6: Event extraction model

and argument predictions for each span in S and predicts the pairing between arguments and triggers to create events from individual span predictions.

Input encoding: Input sentences are mapped to contextualized word embeddings using *Bio+Clinical BERT* (Alsentzer et al., 2019). To limit computational cost, the contextualized word embeddings feed into a bi-LSTM without fine tuning BERT (no backpropagation to BERT). The bi-LSTM has hidden size v_h. The forward and backward states, $\boldsymbol{h}_{t,f}$ and $\boldsymbol{h}_{t,b}$, are concatenated to form the $1 \times 2v_h$ dimensional vector $\boldsymbol{h}_t = [\boldsymbol{h}_{t,f}, \boldsymbol{h}_{t,b}]$, where t is the token position.

Span representation: Each span is represented as the attention weighted sum of the bi-LSTM hidden states. Separate attention mechanisms, c, are implemented for trigger and each labeled argument, and a single attention mechanism is implemented for all span-only arguments, $c \in \{1, 2 \dots 6\}$ (1 for trigger, 4 for labeled arguments, and 1 for span-only arguments). The attention score for span representation c at token position t is calculated as

$$\alpha_{c,t} = \boldsymbol{w}_{\alpha,c}\boldsymbol{h}_t^T \tag{1}$$

where $\boldsymbol{w}_{\alpha,c}$ is a learned $1 \times 2v_h$ vector. For span representation c, span i, and token position t, the attention weights are calculated by normalizing the attention scores as

$$a_{c,i,t} = \frac{\exp(\alpha_{c,t})}{\sum\limits_{k=start(s_i)}^{end(s_i)} \exp(\alpha_{c,k})}, \tag{2}$$

where $start(s_i)$ and $end(s_i)$ denote the start and end token indices of span s_i. Span representation c for span i is calculated as the attention-weighted sum of the bi-LSTM hidden state as

$$\boldsymbol{g}_{c,i} = \sum_{t=start(s_i)}^{end(s_i)} a_{c,i,t}\boldsymbol{h}_t. \tag{3}$$

Span prediction: Similar to the span representations, separate span classifiers, c, are implemented for trigger and each labeled argument, and a single classifier predicts all span-only arguments, $c \in \{1, 2 \dots 6\}$ (1 for trigger, 4 for labeled arguments, and 1 for span-only arguments). Label scores for classifier c and span i are calculated as

$$\phi_c(s_i) = \boldsymbol{w}_{s,c}\text{FFNN}_{s,c}(\boldsymbol{g}_{c,i}), \tag{4}$$

where $\phi_c(s_i)$ yields a vector of label scores of size $|L_c|$, $\text{FFNN}_{s,c}$ is a non-linear projection from size $2v_h$ to v_s, and $\boldsymbol{w}_{s,c}$ has size $|L_c| \times v_s$.

The trigger prediction label set is $L_{trigger} = \{null, \mathit{COVID}, \mathit{Symptom}\}$. Separate classifiers are used for each labeled argument (*Assertion*, *Change*, *Severity*, and *Test Status*) with label set, $L_c = \{null \cup L_l\}$, where L_l is defined in Table 1.[1] For example, $L_{Severity} = \{null, mild, moderate, severe\}$. A single classifier predicts all span-only arguments with label set, $L_{span-only} = \{null, Anatomy, Characteristics, Duration, Frequency\}$.

Argument role prediction: The argument role layer predicts the assignment of arguments to triggers using separate binary classifiers, d, for each labeled argument and one classifier for all span-only arguments, $d \in \{1, 2, \dots 5\}$ (4 for labeled arguments and 1 for span-only arguments). Argument role scores for trigger j and argument k using argument role classifier d are calculated as

$$\psi_d(s_j, s_k) = \boldsymbol{w}_{r,d}\text{FFNN}_{r,d}([\boldsymbol{g}_j, \boldsymbol{g}_k]) \tag{5}$$

where $\psi_d(s_j, s_k)$ is a vector of binary scores of size 2, $\text{FFNN}_{r,d}$ is a non-linear projection from size $2v_s$ to v_r, and $\boldsymbol{w}_{r,d}$ has size $2 \times v_r$.

Span width pruning: To limit time and space complexity of the pairwise argument role predictions, only the top-K spans for each span classifier, c, are considered during argument role prediction. The span score is calculated as the maximum label score in ϕ_c, excluding the $null$ label score.

5 Experimental Setup

5.1 Model Configuration

The model configuration was selected using 3-fold cross validation (CV) on the training set. Table 2 summarizes the selected configuration. Training

[1] The assertion classifier uses the larger label set associateed with *Symptom*.

loss was calculated by summing the cross entropy across all span and argument role classifiers.

Parameter	Value
Maximum sentence length, n	30
Maximum span length, M	6
Top-K spans per classifier	n
Batch size	100
Number of epochs	100
Learning rate	0.001
Optimizer	Adam
Maximum gradient L2-norm	100
BERT embedding dropout	0.3
bi-LSTM hidden size, v_h	200
bi-LSTM activation function	tanh
bi-LSTM dropout	0.3
Span classifier projection size, v_s	100
Span classifier activation function	ReLU
Span classifier dropout	0.3
Role classifier projection size, v_r	100
Role classifier activation function	ReLU
Role classifier dropout	0.3

Table 2: Model configuration

5.2 Data Representation

During initial experimentation, *Symptom Assertion* extraction performance was high for the *absent* subtype and lower for *present*. The higher *absent* performance is primarily associated with the consistent presence of negation cues, like "denies" or "no." While there are affirming cues, like "reports" or "has," the *present* subtype is often implied by a lack of negation cues. For example, an entire sentence could be "Short of breath." To provide the *Symptom Assertion* span classifier with a more consistent span representation, we substituted the *Symptom* trigger token indices for the *Symptom Assertion* token indices in each event and found that performance improved. We extended this trigger token indices substitution approach to all labeled arguments (*Assertion*, *Change*, *Severity*, and *Test Status*) and found performance improved. By substituting the trigger indices for the labeled argument indices, trigger and labeled argument prediction is roughly treated as a multi-label classification problem, although the model is not constrained to require trigger and labeled argument predictions to be paired with the same spans. As previously discussed, the scoring routine does not consider the span indices of labeled arguments.

6 Results

Table 3 presents the extraction performance on the training set using CV and withheld test set. Extraction performance is similar on the train and test sets, even though the training set has higher rates of COVID-19 positive notes. *COVID* trigger extraction performance is very high (0.97 F1) and exceeds the *round 2* annotator agreement (0.95 F1). The *COVID Assertion* performance (0.73 F1) is higher than *Test Status* performance (0.62 F1), which is likely due to the more consistent *Assertion* annotation. *Symptom* trigger and *Assertion* extraction performance is high (0.83 F1 and 0.79 F1, respectively), approaching the *round 2* annotator agreement (0.86 F1 and 0.83 F1, respectively). *Anatomy* extraction performance (0.61 F1) is lower than expected, given the high *round 2* annotator agreement (0.81 F1). *Duration* extraction performance is comparable to annotator agreement, and *Frequency* extraction performance is lower than annotation agreement. *Change*, *Severity*, and *Characteristics* extraction performance is low, again likely related to low annotator agreement for these cases.

As a preliminary analysis, the extraction model was applied to all of the notes in the 92K note data set with a *positive* or *negative COVID-19 Test* label derived from the EHR (1.1K *positive* and 8.0K *negative* notes). Figure 7 presents the pointwise mutual information (PMI) between the EHR derived labels (*positive* or *negative*) and the automatically extracted symptoms with *Assertion* subtypes. Symptoms with *present* and *absent* subtypes are treated as separate features in this figure. The extracted symptoms were manually normalized to aggregate different extracted spans with similar meanings (e.g. "sob"→ "shortness of breath" or "fatigued" → "fatigue"). Only symptoms that occur in at least 3% of notes are included. These results affirm several of the known COVID-19 symptoms, including fever, cough, and shortness of breath. They also suggest that other symptoms, including weakness, diarrhea, vomiting, and myalgia, are indicative of COVID-19 infection. Congestion, anxiety, sore throat, wheezing, and swelling appear to be counter indicators for COVID-19 infection.

This preliminary analysis was performed on a very biased data set. All notes come from a period when testing was very limited, and only patients with prominent COVID-19 symptoms were tested. Additionally, we did not include any patient demographics or past medical history in the analysis. In ongoing COVID-19 work, the extractor will be applied to a broader set of over 2 million ambulatory care and emergency department notes created at

Event type	Argument	Train-CV				Test			
		# Gold	P	R	F1	# Gold	P	R	F1
COVID	Trigger	3,931	0.95	0.97	0.96	1,497	0.96	0.97	0.97
	Assertion	2,936	0.70	0.74	0.72	1,075	0.72	0.74	0.73
	Test Status	1,068	0.60	0.62	0.61	457	0.63	0.60	0.62
Symptom	Trigger	13,823	0.82	0.85	0.83	5,789	0.81	0.85	0.83
	Assertion	13,833	0.77	0.79	0.78	5,791	0.77	0.80	0.79
	Change	739	0.45	0.03	0.06	341	0.45	0.05	0.09
	Severity	743	0.47	0.30	0.37	327	0.45	0.31	0.37
	Anatomy	3,839	0.76	0.59	0.66	1,959	0.78	0.50	0.61
	Characteristics	3,145	0.59	0.26	0.36	1,441	0.66	0.25	0.36
	Duration	3,744	0.62	0.44	0.51	1,344	0.54	0.56	0.55
	Frequency	801	0.64	0.39	0.48	250	0.60	0.51	0.55

Table 3: Extraction performance

UW during the first five months of the pandemic for a more comprehensive analysis of reported symptoms, patient characteristics, and outcomes.

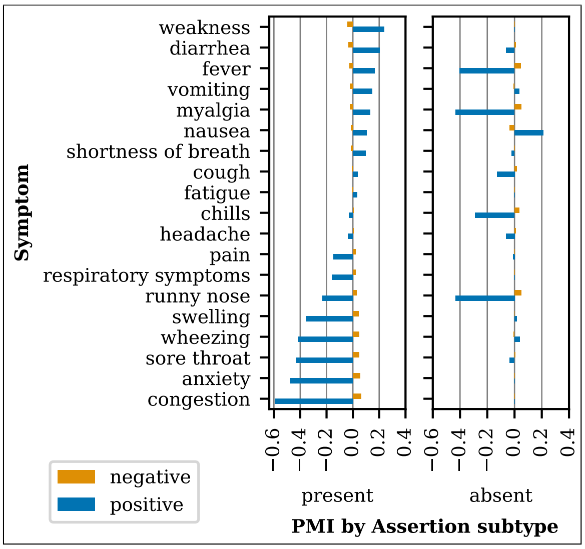

Figure 7: Top automatically extracted symptoms

7 Conclusions

We present CACT, a novel corpus with detailed annotations for COVID-19 diagnoses, testing, and symptoms. CACT includes 1,472 unique notes across six note types with more than 500 notes from patients with positive COVID-19 tests. We implement a span-based event extraction model that jointly extracts all annotated phenomena, including argument types and subtypes. The extraction model performs well in the extraction of *COVID* trigger (0.97 F1) and *Assertion* (0.73 F1) and achieves near-human performance in the extraction of *Symptom* trigger (0.83 F1) and *Assertion* (0.79 F1). The automatic extractor was applied to a pool of 9.1K unannotated notes, providing a preliminary exploration of key COVID-19 symptoms.

In future work, the extractor will be applied to a much larger set of clinical ambulatory care and emergency department notes from UW and collaborating institutions nationally. The extracted symptom information will also be combined with routinely coded data (e.g. diagnosis and procedure codes, demographics) and automatically extracted data (e.g. social determinants of health). Using these data, we will develop models for predicting risk of COVID-19 infection amongst individuals who are tested. These models could better inform clinical indications for prioritizing testing with constrained test availability and more accurately determine pre-test probability. Additionally, the presence or absence of certain symptoms can be used to inform clinical care decisions with greater precision. This future work may also identify combinations of symptoms (including their presence, absence, severity, sequence of appearance, duration, etc.) associated with clinical outcomes and health service utilization, such as deteriorating clinical course and need for repeat consultation or hospital admission. The use of detailed symptom information will be highly valuable in informing these models, but potentially only with the level of nuance that our extraction models provide. With the COVID-19 pandemic continuing for the foreseeable future, accelerating the research outlined in this paper will inform key clinical and health service decision making.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Clinical Natural Language Processing Workshop*, pages 72–78.

Long Chen, Yu Gu, Xin Ji, Zhiyong Sun, Haodan Li, Yuan Gao, and Yang Huang. 2019. Extracting medications and associated adverse drug events using a natural language processing system combining knowledge base and deep learning. *Journal of the American Medical Informatics Association*, 27(1):56–64.

Fenia Christopoulou, Thy Thy Tran, Sunil Kumar Sahu, Makoto Miwa, and Sophia Ananiadou. 2020. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *Journal of the American Medical Informatics Association*, 27(1):39–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Weipeng Huang, Xingyi Cheng, Taifeng Wang, and Wei Chu. 2019. Bert-based multi-head selection for joint entity-relation extraction. In *International Conference on Natural Language Processing and Chinese Computing*, pages 713–723.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, 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*, 3:160035.

Rob Koeling, John Carroll, Rosemary Tate, and Amanda Nicholson. 2011. Annotating a corpus of clinical text records for learning to recognize symptoms automatically. In *International Workshop on Health Text Mining and Information Analysis*, pages 43–50.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Empirical Methods in Natural Language Processing*, pages 188–197.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Empirical Methods in Natural Language Processing*, pages 3219–3232.

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. In *North American Chapter of the Association for Computational Linguistics*, pages 3036–3046.

Walker Orr, Prasad Tadepalli, and Xiaoli Fern. 2018. Event detection with neural networks: A rigorous empirical evaluation. In *Conference on Empirical Methods in Natural Language Processing*, pages 999–1004.

Yihe Pang, Jie Liu, Lizhen Liu, Zhengtao Yu, and Kai Zhang. 2019. A deep neural network model for joint entity and relation extraction. *IEEE Access*, 7:179143–179150.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *North American Chapter of the Association for Computational Linguistics*, pages 2227–2237.

Guoqing Qian, Naibin Yang, Ada Hoi Yan Ma, Liping Wang, Guoxiang Li, Xueqin Chen, and Xiaomin Chen. 2020. COVID-19 transmission within a family cluster by presymptomatic carriers in china. *Clinical Infectious Diseases*.

Hagai Rossman, Ayya Keshet, Smadar Shilo, Amir Gavrieli, Tal Bauman, Ori Cohen, Esti Shelly, Ran Balicer, Benjamin Geiger, Yuval Dor, et al. 2020. A framework for identifying regional outbreak and spread of covid-19 from one-minute population-wide surveys. *Nature Medicine*, pages 1–4.

Xue Shi, Yingping Yi, Ying Xiong, Buzhou Tang, Qingcai Chen, Xiaolong Wang, Zongcheng Ji, Yaoyun Zhang, and Hua Xu. 2019. Extracting entities with attributes in clinical text via joint deep learning. *Journal of the American Medical Informatics Association*, 26(12):1584–1591.

Brett R South, Shuying Shen, Makoto Jones, Jennifer Garvin, Matthew H Samore, Wendy W Chapman, and Adi V Gundlapalli. 2009. Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. In *BMC Bioinformatics*, volume 10.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

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*, 18(5):552–556.

Pauline Vetter, Diem Lan Vu, Arnaud G L'Huillier, Manuel Schibler, Laurent Kaiser, and Frederique Jacquerioz. 2020. Clinical features of covid-19. *British Medical Journal*.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations.

In *Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 5788–5793.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus ldc2006t06.

Haoyu Wang, Ming Tan, Mo Yu, Shiyu Chang, Dakuo Wang, Kun Xu, Xiaoxiao Guo, and Saloni Potdar. 2019. Extracting multiple-relations in one-pass with pre-trained transformers. In *Association for Computational Linguistics*, pages 1371–1377.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Michael Kinney, Ziyang Liu, William. Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020a. Cord-19: The covid-19 open research dataset. *arXiv*.

Xuan Wang, Xiangchen Song, Yingjun Guan, Bangzheng Li, and Jiawei Han. 2020b. Comprehensive named entity recognition on cord-19 with distant or weak supervision. *arXiv*.

Wycliffe E Wei, Zongbin Li, Calvin J Chiew, Sarah E Yong, Matthias P Toh, and Vernon J Lee. 2020. Presymptomatic transmission of SARS-CoV-2—singapore, january 23–march 16, 2020. *Morbidity and Mortality Weekly Report*, 69(14):411.

World Health Organization. 2020a. Coronavirus disease (covid-19) situation report – 154.

World Health Organization. 2020b. Global literature on coronavirus disease.

Chaomin Wu, Xiaoyan Chen, Yanping Cai, Xing Zhou, Sha Xu, Hanping Huang, Li Zhang, Xia Zhou, Chunling Du, Yuye Zhang, et al. 2020. Risk factors associated with acute respiratory distress syndrome and death in patients with coronavirus disease 2019 pneumonia in wuhan, china. *JAMA Internal Medicine*.

Zunyou Wu and Jennifer M. McGoogan. 2020. Characteristics of and important lessons from the coronavirus disease 2019 (covid-19) outbreak in china: Summary of a report of 72314 cases from the chinese center for disease control and prevention. *Journal of the American Medical Association*, 323(13):1239–1242.

Jing Yang, Ya Zheng, Xi Gou, Ke Pu, Zhaofeng Chen, Qinghong Guo, Rui Ji, Haojia Wang, Yuping Wang, and Yongning Zhou. 2020. Prevalence of comorbidities in the novel wuhan coronavirus (covid-19) infection: a systematic review and meta-analysis. *International Journal of Infectious Diseases*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Suncong Zheng, Yuexing Hao, Dongyuan Lu, Hongyun Bao, Jiaming Xu, Hongwei Hao, and Bo Xu. 2017. Joint entity and relation extraction based on a hybrid neural network. *Neurocomputing*, 257:59 – 66. Machine Learning and Signal Processing for Big Multimedia Analysis.