JAMIA Open, 4(3), 2021, 1–9
doi: 10.1093/jamiaopen/ooab070
Research and Applications

Research and Applications

A fast, resource efficient, and reliable rule-based system for COVID-19 symptom identification

Himanshu S Sahoo [1,2,†], Greg M Silverman [2,†], Nicholas E Ingraham [3], Monica I Lupei [4], Michael A Puskarich [5], Raymond L Finzel [6], John Sartori [1], Rui Zhang [6,7], Benjamin C Knoll [7], Sijia Liu [8], Hongfang Liu [8], Genevieve B Melton [2,7], Christopher J Tignanelli [2,‡] and Serguei V S Pakhomov [6,‡]

[1]Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, Minnesota, USA, [2]Department of Surgery, University of Minnesota, Minneapolis, Minnesota, USA, [3]Pulmonary Disease and Critical Care Medicine, University of Minnesota, Minneapolis, Minnesota, USA, [4]Department of Anesthesiology, University of Minnesota, Minneapolis, Minnesota, USA, [5]Department of Emergency Medicine, University of Minnesota, Minneapolis, Minnesota, USA, [6]Department of Pharmaceutical Care and Health Systems, University of Minnesota, Minneapolis, Minnesota, USA, [7]Institute for Health Informatics, University of Minnesota, Minneapolis, Minnesota, USA and [8]Department of Health Science Research, Mayo Clinic, Rochester, New York, USA

[†]These authors contributed equally to this work as first authors.
[‡]These authors contributed equally to this work as senior authors.
Corresponding Author: Himanshu S. Sahoo, MS, Department of Electrical and Computer Engineering, University of Minnesota, KH 4-166, Kenneth H. Keller Hall, 200 Union St SE, Minneapolis, MN, USA (sahoo009@umn.edu)

Received 16 March 2021; Revised 16 July 2021; Editorial Decision 30 July 2021; Accepted 5 August 2021

ABSTRACT

Objective: With COVID-19, there was a need for a rapidly scalable annotation system that facilitated real-time integration with clinical decision support systems (CDS). Current annotation systems suffer from a high-resource utilization and poor scalability limiting real-world integration with CDS. A potential solution to mitigate these issues is to use the rule-based gazetteer developed at our institution.
Materials and Methods: Performance, resource utilization, and runtime of the rule-based gazetteer were compared with five annotation systems: BioMedICUS, cTAKES, MetaMap, CLAMP, and MedTagger.
Results: This rule-based gazetteer was the fastest, had a low resource footprint, and similar performance for weighted microaverage and macroaverage measures of precision, recall, and f1-score compared to other annotation systems.
Discussion: Opportunities to increase its performance include fine-tuning lexical rules for symptom identification. Additionally, it could run on multiple compute nodes for faster runtime.
Conclusion: This rule-based gazetteer overcame key technical limitations facilitating real-time symptomatology identification for COVID-19 and integration of unstructured data elements into our CDS. It is ideal for large-scale deployment across a wide variety of healthcare settings for surveillance of acute COVID-19 symptoms for integration into prognostic modeling. Such a system is currently being leveraged for monitoring of postacute sequelae of COVID-19 (PASC) progression in COVID-19 survivors. This study conducted the first in-depth analysis and developed a rule-based gazetteer for COVID-19 symptom extraction with the following key features: low processor and memory utilization, faster runtime, and similar weighted microaverage and macroaverage measures for precision, recall, and f1-score compared to industry-standard annotation systems.

LAY SUMMARY

With COVID-19 came an unprecedented need to identify symptoms of COVID-19 patients under investigation (PUIs) in a time-sensitive, resource-efficient, and accurate manner. While available annotation systems perform well for smaller healthcare settings, they fail to scale in larger healthcare systems where 10 000+ clinical notes are generated a day. This study covers three improvements addressing key limitations of current annotation systems. (1) High resource utilization and poor scalability of existing annotation systems. The presented rule-based gazetteer is a high-throughput annotation system for processing high volume of notes, thus, providing an opportunity for clinicians to make more informed time-sensitive decisions around patient care. (2) Equally important is our developed rule-based gazetteer performs similar or better than current annotation systems for symptom identification. (3) Due to the minimal resource needs of the rule-based gazetteer, it could be deployed at healthcare sites lacking a robust infrastructure where industry-standard annotation systems cannot be deployed because of low resource availability.

Key words: natural language processing, clinical decision support systems, artificial intelligence, information extraction, signs, and symptoms, follow-up studies

BACKGROUND AND SIGNIFICANCE

With COVID-19 came an unprecedented need to identify symptoms of COVID-19 PUIs in a time-sensitive, resource-efficient, and accurate manner. When attempting to identify COVID-19 symptoms from clinical notes in near-real-time, we identified significant limitations with industry-standard annotation systems (hereby referred to as "annotation systems") including (1) poor scalability with increasing number of notes and (2) high resource needs.

While available annotation systems perform well for smaller healthcare settings, they fail to scale in larger healthcare systems (like ours), where 10 000+ clinical notes are generated a day. For example, one instance of MetaMap takes approximately 105 h; CLAMP 28 h, and cTAKES 9 h to process 12 000 notes limiting scalability especially for time-sensitive prognosis such as for COVID-19 PUIs. Similar issues were also found by other researchers.[1,2] Solutions proposed to mitigate scalability issues included: increasing number of servers, NLP engines, and databases. Although these solutions led to improved runtime, they still did not address the key issue of high resource utilization, being problematic for healthcare sites lacking robust infrastructure.

After evaluating several potential annotation systems to address the above-mentioned limitations, we developed a solution using a dictionary of terms (called as a gazetteer) with significantly lower resource utilization, faster runtime, and similar weighted microaverage and macroaverage measures compared to annotation systems. When time-sensitive decisions with minimal patient contact are crucial, such as during the COVID-19 pandemic, this was extremely important. This study presents our findings.

Multiple studies have demonstrated the success of rule-based gazetteers consisting of domain-specific lexica as an alternative to annotation systems. In one study, Liu et al.[3] successfully used a gazetteer to select cohorts of heart failure and peripheral arterial disease patients from unstructured text, while Wagholikar et al.[4] used a gazetteer based on radiological findings to automate limb fracture classification. Gazetteer lexicons are highly targeted within clinical domains through construction by subject matter experts, especially when combined with appropriate lexical rules[5,6] and work very well[7] with continuous maintenance.[8] Gazetteers can easily be deployed together as a standalone tool using containerization technologies, and their rule-base alone can be deployed as part of an existing infrastructure, such as developed by the Open Health NLP (OHNLP) consortium for the National COVID Cohort Collaborative (N3C).[9,10]

This study developed a rule-based gazetteer based on a lexicon of COVID-19 symptoms (hereby referred to as "COVID-19 gazetteer") and compared it to five annotation systems in terms of (1) document processing times; (2) resource needs; and (3) performance in terms of weighted microaverage and macroaverage measures for precision, recall, and f1-score.

MATERIALS AND METHODS

Metrics used for comparing annotation systems

Runtime

Amount of time taken by an annotation system to process a given set of documents.

Resource utilization

Central processing units (CPUs) and random access memory (RAM) utilized by an annotation system. Henceforth, CPUs are referred to as "processor" and RAM is referred to as "memory."

Weighted microaverage and macroaverage measures

Weighted microaverage (henceforth referred to as "microaverage performance measures") and macroaverage measures for positive predictive value (precision), sensitivity (recall), and harmonic mean (f1-score) for the task of symptom identification.

System overview

Runtime evaluations were performed on a computing system with configurations listed in Supplementary Appendix A. All annotation systems were containerized using Docker.[11] To ensure equal access to system resources all tests were serially executed in a Kubernetes/Argo[12] workflow where each annotation system ran as a single Kubernetes pod.

Data

Notes were collected from M Health Fairview affiliated with the University of Minnesota (UMN), comprising 12 hospitals and services in the ambulatory and postacute settings. There are over 368 000 ED visits with 9% to 30% cases admitted as inpatients each year. Between March 2020 to December 2020 there were 19 924 total ED visits for 10 110 unique COVID-19 positive patients. 12 000

notes were randomly selected from the pool of ED notes for comparing runtime and resource utilization of the annotation systems.

Expert-curated manually annotated reference corpora

At the time of this study, there were no existing corpora annotated for COVID-19 symptoms. Small corpora of notes were quickly developed by UMN and Mayo Clinic to assess COVID-19 symptom identification performance of annotation systems. Due to the small corpora size, the results obtained would not be sufficient to establish which annotation system is better than the other for symptom identification. In their study of UMLS-extracted symptoms for predictive modeling of influenza,[13] Stephens et al. used only 20 randomly selected notes (with only 200 labeled symptoms) to assess their extraction process, suggesting the corpora used in this study is adequate for testing symptom identification performance and finding potential gaps between annotation systems.

UMN reference corpus

Forty-six notes from M Health Fairview (hereby referred to as "UMNCor") were randomly selected and manually annotated by a board-certified critical care physician with 12 years of clinical experience who is also a board-certified clinical informaticist (CT). The annotator had experience treating over 250 COVID-19 positive patients and was blinded to the results of annotation systems. Notes in UMNCor were manually reviewed for positive and negative document-level mentions of 11 acute COVID-19 symptoms as documented by the Center for Disease Control and Prevention (CDC)[14] (hereby referred to as "acute CDC symptoms"). The phrase "positive document-level mention" means at least one positive mention of the acute CDC symptom in the entire note. Similarly, the phrase "negative document-level mention" means at least one negative mention. UMNCor contained a total of 259 document-level mentions (shown in Table 1).

Mayo reference corpus

This corpus, developed by Mayo Clinic, consists of 148 fully deidentified notes for COVID-19 positive patients (hereby referred to as "MayoCor"). Each note was labeled for symptoms based on the CDC and Mayo lexicons.[14,15] The annotation guidelines were developed in collaboration with the Coronavirus Infectious Disease Ontology (CIDO) team.[16] MayoCor contained a total of 260 document-level mentions (shown in Table 1).

Symptom selection criteria

Only acute CDC symptoms were included in the study. Any document-level mention with negligible number of instances compared to the mention with the highest number of instances will not contribute much to microaverage performance measures. Hence, document-level mentions with less than five instances were excluded for the calculation of microaverage performance measures. Using the above-mentioned criteria and Table 1, document-level mentions included for calculation of microaverage performance measures for both corpora are mentioned in Supplementary Appendix C. These mentions selected for UMNCor and MayoCor for microaverage performance measures calculations are hereby referred to as "UMNCor features" and "MayoCor features," respectively.

For macroaverage measures of precision, recall, and f1-score, positive and negative document-level mentions of all acute CDC symptoms have been used for calculation. Since macroaverage measures assign equal weight to every class it is worthwhile to examine how annotation systems compare when treating every acute CDC symptom equally irrespective of sample size.

Table 1. Count of document-level mentions for acute CDC symptoms for the corpora

Features	No. of mentions	
	UMNCor	MayoCor
cdc_aches_n	3	3
cdc_aches_p	18	3
cdc_cough_n	11	6
cdc_cough_p	28	22
cdc_diarrhea_n	12	14
cdc_diarrhea_p	11	18
cdc_dyspnea_n	10	15
cdc_dyspnea_p	28	34
cdc_fatigue_n	2	1
cdc_fatigue_p	15	14
cdc_fever_n	9	24
cdc_fever_p	30	18
cdc_headaches_n	5	8
cdc_headaches_p	8	15
cdc_nausea_vomiting_n	19	20
cdc_nausea_vomiting_p	13	27
cdc_rhinitis_congestion_n	7	1
cdc_rhinitis_congestion_p	8	2
cdc_sore_throat_n	6	4
cdc_sore_throat_p	9	3
cdc_taste_smell_loss_n	2	3
cdc_taste_smell_loss_p	5	5
sum	259	260

Note: suffix "_p" following an acute CDC symptom represents positive document-level mention for the acute CDC symptom and while suffix "_n" represents negative document-level mention

Lexicon of COVID-19 symptoms

Lexicon of 171 terms based on the CDC's guidelines was iteratively created by three board-certified clinicians (NI, ML, and MP), using equivalent medical terminology, abbreviations, synonyms, allied symptoms, alternate spellings, misspellings, etc. Terms in this lexicon (see Supplementary Appendix B.1) hereafter referred to as "Derived COVID-19 Symptoms" were used by the COVID-19 gazetteer and to derive the Universal Medical Language System (UMLS)[17] lexicon used by other annotation systems.

Query expansion of derived COVID-19 symptoms

We utilized *word2vec* model[18] trained on clinical text by Pakhomov et al.[19] to expand the derived COVID-19 symptoms list (see Supplementary Appendix B.2). The model was trained on a corpus of notes (4 169 696 714 tokens) from M Health Fairview between 2010 to 2014, inclusive. The model created embeddings with up to four-word sequences by using the word2phrase tool.[18] The 2018 version of MetaMap was used to map lexicon terms to the UMLS. The final set of terms mapped to UMLS concepts was further reviewed by three board-certified clinicians (NI, ML, and MP) to ensure semantic expansions were clustered appropriately on the acute CDC symptoms. This final set of terms and concepts (see Supplementary Appendix B.4) was made available as a UMLS lexicon for use by annotation systems (refer to subsection "UIMA-based annotation pipeline").

UIMA-based annotation pipeline

Notes were annotated for Concept Unique Identifiers (CUIs) from the *Disorders* semantic group using the NLP Artifact Discovery and Preparation Toolkit for Kubernetes (NLP-ADAPT-kube).[20] NLP-ADAPT-kube contains the following Unstructured Information Management Architecture (UIMA)[21] compatible annotation systems as Docker images: (1) BioMedICUS v2.2.0[22]; (2) cTAKES v4.0.1[23]; (3) MetaMap 2018 Linux version[24]; (4) CLAMP v1.6.4.[25] Features relevant to the acute CDC symptoms were constructed using extracted UMLS concepts present in the derived UMLS lexicon described in subsection "Query expansion of derived COVID-19 symptoms."

MedTagger

MedTagger v1.0.9[26] is a rule-based gazetteer developed by the Mayo Clinic. We used two versions of MedTagger: (1) COVID-19 gazetteer lexicon adopted to MedTagger's ruleset format (hereby referred to as "MedTagger Custom") and (2) Mayo Clinic's COVID-19 lexicon[27] adopted to MedTagger's ruleset format.

COVID-19 gazetteer

The COVID-19 gazetteer used the lexicon described in subsection "Lexicon of COVID-19 symptoms" to narrow searches for concepts belonging to sentence-level mentions of acute CDC symptoms. The COVID-19 gazetteer uses *spaCy's* Matcher[28] and EntityRuler[29] classes to add lexicon terms to the *spaCy* **en_core_web_sm**[30] model. The Matcher instance reads in notes and returns the span of text containing symptom mentions. Returned spans are further processed by the *spaCy* pipeline to search for custom entities added using EntityRuler. This extra step is necessary because we observed *spaCy* missed certain phrases in the lexicon; thus, the Matcher instance detected terms the EntityRuler instance had missed. Span length was predetermined through initial tuning on a held-out set of 1700 randomly selected notes. Output was then lemmatized to convert text to its base form (eg, the base form of "wheezing" is "wheeze"). The NegEx component of *spaCy* (*negspaCy*[31]) was added at the end of the *spaCy* pipeline for negation detection. More details about the COVID-19 gazetteer are present in the GitHub repository.[32] The COVID-19 gazetteer used multiple server cores by distributing nearly equal numbers of notes to each core.

RESULTS

Overall microaverage performance measures of annotation systems are shown for both corpora in Table 2. As mentioned in subsection "Symptom selection criteria," UMNCor uses only UMNCor features and MayoCor uses only MayoCor features for calculating microaverage performance measures.

Table 3 shows the macroaverage measures for precision, recall, and f1-score for positive and negative document-level mentions for all the acute CDC symptoms (as mentioned in subsection "Symptom selection criteria").

Figure 1 shows total CPU and RAM utilization for the annotation systems over their runtime on 9000 clinical notes. Total utilization values for CPU and RAM (referred to as cores*sec and RAM*sec in Figure 1, respectively) were calculated as a running summation of the CPU (in cores) and RAM (in gigabytes (GB)) utilized by an annotation system over its runtime. The ideal system would minimize resources while executing in the least amount of time. MedTagger Custom was omitted from any runtime analysis because it uses the same underlying implementation as MedTagger.

Figure 2 shows the runtimes of the annotation systems when run on 9000 clinical notes.

To demonstrate the efficiency of the COVID-19 gazetteer we analyzed its runtime by keeping the number of notes constant while increasing the number of cores for 3000, 6000, 9000, and 12 000 clinical notes (see Figure 3).

Table 2. Overall microaverage performance measures of the annotation systems for both corpora (confidence intervals are present in Supplementary Appendix D.1-2)

	UMNCor			MayoCor		
System	Precision	Recall	f1-score	Precision	Recall	f1-score
BioMedICUS	0.78	0.75	0.75	0.89	0.89	0.89
CLAMP	0.84	0.85	0.85	0.91	0.92	0.91
cTAKES	0.83	0.80	0.81	0.91	0.90	0.91
MetaMap	0.85	0.84	0.85	0.90	0.91	0.90
COVID-19 Gazetteer	0.89	0.86	0.87	0.91	0.91	0.91
MedTagger Custom	0.82	0.82	0.82	0.92	0.92	0.92
MedTagger	0.88	0.85	0.85	0.91	0.91	0.91

DISCUSSION

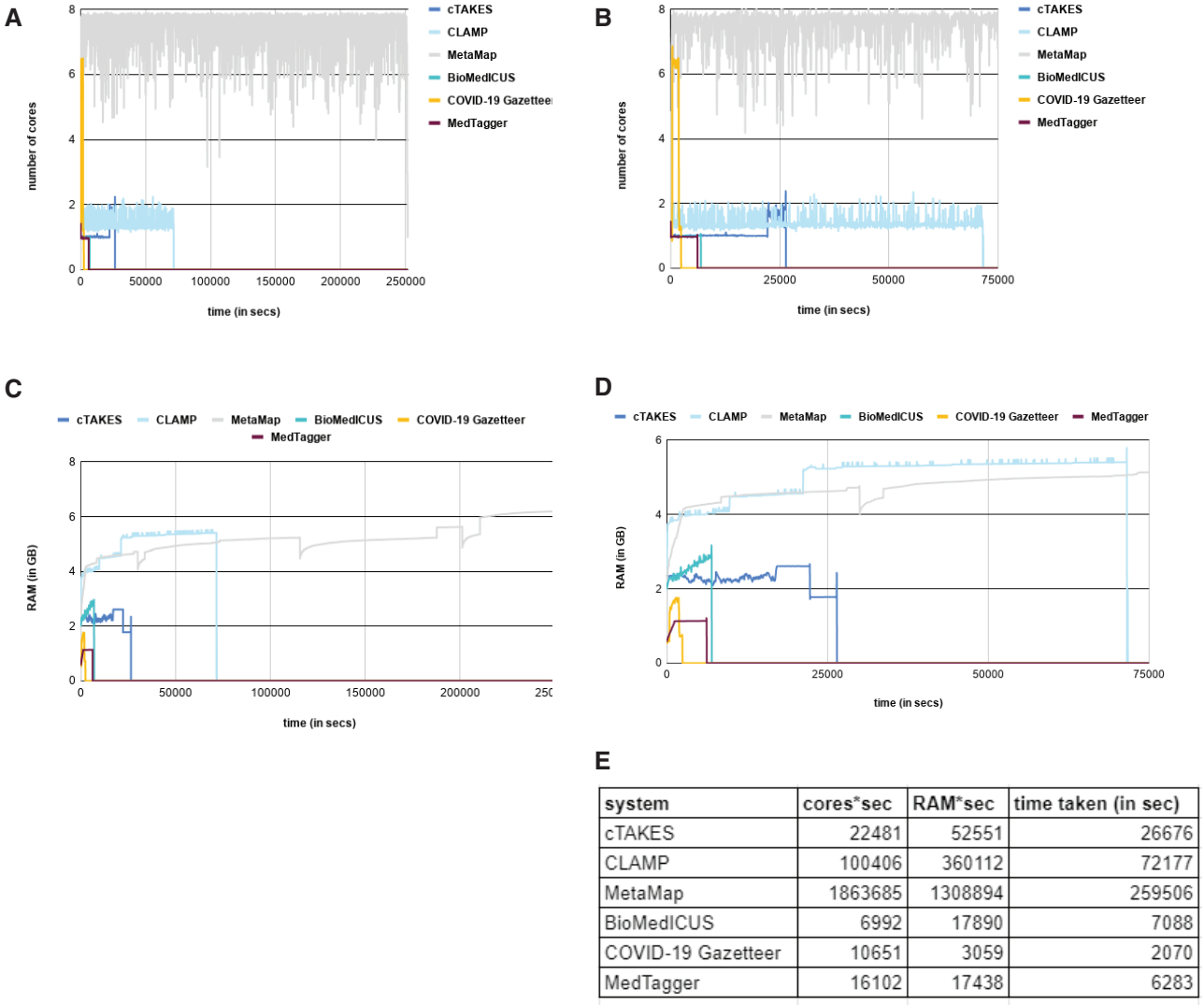
The purpose of this study was to develop a rule-based gazetteer for COVID-19 and compare it to five annotation systems. This study makes the following contributions: (1) first in-depth analysis involving rule-based gazetteer for COVID-19 symptom identification; (2) compares performance (weighted microaverage and macroaverage measures for precision, recall, and f1-score) of the COVID-19 gazetteer to other annotation systems; (3) highlights the potential of the COVID-19 gazetteer as a low resource solution by comparing its processor and memory utilization to other annotation systems; (d) compares runtime of the COVID-19 gazetteer to other annotation systems, demonstrating its efficacy for high-throughput real-time annotation of notes for identifying a patient's presenting symptoms[33] (eg, identifying symptoms of COVID-19 PUIs in a time-sensitive manner).

Performance of systems

Results in Tables 2 and 3 and Supplementary Appendices D and E demonstrate the COVID-19 gazetteer has similar weighted microa-

Table 3. Macroaverage performance measures of the annotation systems for both corpora calculated using positive and negative document-level mentions for all the acute CDC symptoms (confidence intervals are present in Supplementary Appendix E.1-2)

	UMNCor			MayoCor		
System	Precision	Recall	f1-score	Precision	Recall	f1-score
BioMedICUS	0.71	0.75	0.72	0.73	0.74	0.73
CLAMP	0.81	0.81	0.81	0.79	0.71	0.74
cTAKES	0.77	0.82	0.79	0.75	0.78	0.76
MetaMap	0.80	0.82	0.81	0.75	0.71	0.73
COVID-19 Gazetteer	0.82	0.88	0.84	0.77	0.79	0.78
MedTagger Custom	0.77	0.78	0.77	0.79	0.80	0.80
MedTagger	0.80	0.87	0.82	0.80	0.75	0.77



E

system	cores*sec	RAM*sec	time taken (in sec)
cTAKES	22481	52551	26676
CLAMP	100406	360112	72177
MetaMap	1863685	1308894	259506
BioMedICUS	6992	17890	7088
COVID-19 Gazetteer	10651	3059	2070
MedTagger	16102	17438	6283

Figure 1. Total CPU and RAM utilization over the period of execution of the annotation systems on 9000 notes. A, CPU utilization (in number of cores); B, Zoomed in view of (A); C, RAM utilization; D, Zoomed in view of (C); E, Total utilization of CPU (represented as cores*s) and RAM (represented as RAM*s). Statistics for CPU and RAM utilization were collected every 30 s and appended to a file using a bash script that queried the Kubernetes cluster.

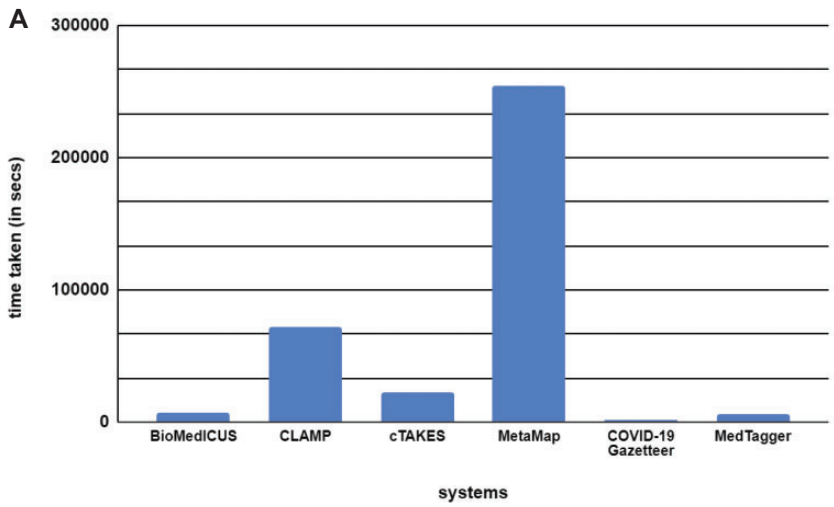


Figure 2. Runtime of annotators for 9000 notes. The COVID-19 gazetteer had the least processing time.

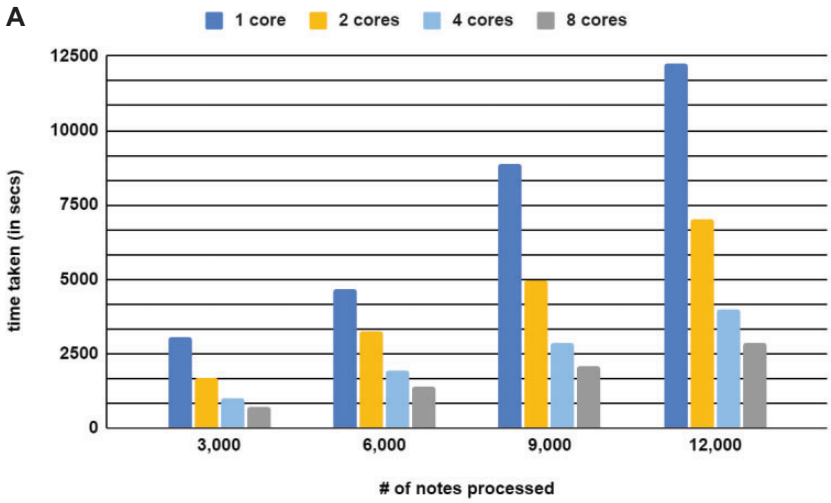


Figure 3. Runtime of the COVID-19 gazetteer with increasing number of CPU cores on a given set of notes. It is observed that runtime reduced as number of cores increased for constant set of notes processed.

verage and macroaverage performance measures compared to other annotation systems. Based on these results, we emphasize the importance of a carefully designed gazetteer for diseases with manageable sets of defined symptoms translatable to lexical rules to aid CDS, including surveillance for long-term care[34] (eg, PASC progression in COVID-19 survivors).

Resource utilization of annotation systems

Figure 1 demonstrates that BioMedICUS, COVID-19 gazetteer, and MedTagger had the lowest CPU and RAM utilization making them good candidates for compute devices with minimal processor and memory resources compared to MetaMap and CLAMP (had highest resource requirements). BioMedICUS utilizes a fast algorithm along with in-memory maps for concept detection but comes with the tradeoff of increased memory utilization. The COVID-19 gazetteer uses **en_core_sci_sm** and **en_core_web_sm** *spaCy* models (about 13–15 MB) for detection of mentions. This is one possible reason why the COVID-gazetteer had the lowest memory requirement. The COVID-19 gazetteer used all available cores to minimize runtime and was among the lowest in terms of overall CPU utilization although the average CPU utilized at any given instant of time was high. MedTagger had low CPU utilization because it processes documents through data streams and loads the compiled ruleset to memory for lower memory utilization. It should be noted annotation systems with minimal resource requirements (eg, BioMedICUS, COVID-19 gazetteer, and MedTagger) have the potential to incur a significantly lower monetary costs when run on cloud-based platforms. In addition, annotation systems with minimal resource requirements are ideal for deployment at healthcare sites lacking robust infrastructure.

Scaling of annotation systems for real-time processing of notes

Results in Figure 2 show the COVID-19 gazetteer consistently outperformed other annotation systems in runtime. The COVID-19 gazetteer took 34 min to process 9000 notes—about 3× faster than MedTagger (second fastest annotation system) and 123× faster than MetaMap (slowest annotation system). Hence, the COVID-19 gazetteer is the best candidate for high-throughput real-time processing of notes for clinical surveillance (eg, identifying symptoms of COVID-19 PUIs). Figure 3 shows the effect of scaling the COVID-19 gazetteer through increase of CPU cores on a given set of notes, where runtime decreases linearly with increasing cores. The COVID-19 gazetteer operating on multiple compute nodes has far greater potential to significantly decrease the runtime to process notes compared to standard annotation systems.

It is possible to scale "off-the-shelf" annotation systems for real-time processing through both pipeline customization[35] and across multiple compute nodes. Demner-Fushman et al. introduced MetaMap Lite[36] and found it to be at least 5× faster than MetaMap and cTAKES on various corpora[37–40] with higher precision, recall, and f1-score. Stephens *et al.* used MetaMap Lite for processing speed and ease of use and compared it to MetaMap and cTAKES on a corpus containing 7278 EHR notes.[13] In the workshop on 'Large Scale Ensembled NLP Systems' with Docker and Kubernetes[41] Finzel et al. scaled MetaMap by running 80 Kubernetes pods on 8 compute nodes to get a processing speed of about 15 documents per second. The study conducted by Miller *et al.* to extract patients' phenotypes from 10 000 EHR notes had a processing speed of about 2.45 notes per second when run on Amazon Web Services (AWS) containing 2 CPUs and 8 GB of RAM.[2] This was equivalent to processing 1 million notes per day when run on 10 AWS Elastic Computing (EC2) nodes. The presentation on 'Fault-Tolerant, Distributed, and Scalable Natural Language Processing with cTAKES'[42] discusses scaling cTAKES using distributed Apache Spark[43] on 245 worker machines, each with 64 CPUs and 240 GBs of RAM. The developed pipeline was able to process 84 387 661 notes in 89 min compared to about 396 days for cTAKES. Despite such high processing capacity, these systems incur incredibly high resource utilization.

Lexicon creation and maintenance for annotation systems

The COVID-19 gazetteer lexicon process (described in subsection "Lexicon of COVID-19 symptoms") required clinical expertise. This process could be automated using transformer models like Bidirectional Encoder Representations from Transformers (BERT).[44] Preliminary experiments conducted in our lab indicate that using only 40 terms representing acute CDC symptoms to fine-tune a

BERT model for Named Entity Recognition (NER) yielded 360 terms belonging to the acute CDC symptoms from 10 000 ED notes for COVID-19 positive patients (refer Supplementary Appendix F for details on BERT setup for NER). BERT symptom extraction process took about 6 h compared to several weeks required by subject matter experts to create the COVID-19 gazetteer lexicon. In addition, the lexicon of 360 terms extracted using BERT had similar symptom identification performance on UMNCor with respect to microaverage performance measures compared to the 171 terms of the COVID-19 gazetteer lexicon created using clinical expertise. As there are variations in lexical constructs while documenting symptoms among medical scribes as well as over time, it is necessary to maintain the COVID-19 gazetteer lexicon by periodically checking for new lexical constructs of acute CDC symptoms present in notes that were not present in the existing the COVID-19 gazetteer lexicon. This could be done by either using the COVID-19 gazetteer lexicon creation process outlined in subsection "Lexicon of COVID-19 symptoms" or by using transformer models like BERT. The COVID-19 gazetteer lexicon could also be easily extended to COVID-19 symptoms not present in the list of acute CDC symptoms.[33] This process would also work for any disease with a well-defined symptomatology, including PASC.

On the other hand, UMLS lexicon creation for UIMA-based annotation systems required the steps mentioned in subsection "Query expansion of derived COVID-19 symptoms" in addition to the COVID-19 gazetteer lexicon creation process. Maintenance of the UMLS lexicon also requires periodically searching for new lexical constructs of acute CDC symptoms present in clinical notes and mapping them to UMLS concepts using rules used to create the UMLS lexicon. Mapping of new lexical constructs to UMLS concepts cannot be automated. This requires costly subject matter intervention and is time-consuming.

To summarize, the UMLS lexicon creation process took two steps compared to a single step required for creating COVID-19 gazetteer lexicon. In addition, the second step of UMLS lexicon creation required extensive clinical expertise. Hence, the COVID-19 gazetteer lexicon was simpler to create and maintain compared to the UMLS lexicon.

Complementing UMLS with gazetteer lexicon

Results in Tables 2 and 3 and Supplementary Appendices D and E confirm the COVID-19 gazetteer performs similar to any annotation system reliant on the UMLS Metathesaurus. The COVID-19 gazetteer lexicon consisted of 120 UMLS terms out of 171 terms. For these 120 UMLS terms, we observed a weighted microaverage f1-score of 0.85 across all the mentions present in UMNCor features, which is 2% less than the observed overall microaverage f1-score of 0.87 for the COVID-19 gazetteer (shown in Table 2). With the use of the remaining 51 non-UMLS terms, the COVID-19 gazetteer improved the matching of relevant terms not detected by the 120 UMLS terms. Thus, the COVID-19 gazetteer lexicon complements the UMLS lexicon making it an ideal candidate for being a part of an ensemble of different UIMA-based annotation systems reliant on a UMLS lexicon. Use of a non-UMLS rule-based gazetteer complemented by UMLS terms could be tailored to any disease with a clearly defined symptomatology.

Limitations and future work

Corpora used consisted of a limited number of document-level mentions—259 for UMNCor and 260 for MayoCor. Due to the small corpora size, the annotation systems had mostly wide and overlapping confidence intervals for weighted microaverage and macroaverage performance measures of precision, recall, and f1-score. Thus, the small corpora size failed to highlight the differences between the annotation systems. However, in their study of UMLS-extracted symptoms for predictive modeling of influenza,[13] Stephens et al. used only 20 randomly selected notes (with only 200 labeled symptoms) to assess their extraction process, suggesting the corpora of notes used in this study is adequate for testing. To address the issue of generalizability and assess significant differences between annotation systems for the task of symptom identification, we are in the process of creating a larger reference corpus of notes manually annotated by multiple raters.

To understand some of the limitations of the COVID-19 gazetteer for future improvements, we manually audited the output of the gazetteer against a few notes from UMNCor. We observed the span of text containing the mention of an acute CDC symptom analyzed by the COVID-19 gazetteer was sometimes too short to contain the negation for the mention. For example, the COVID-19 gazetteer detected a positive mention instead of a negative mention for "sore throat" in the following sentence:

> "The patient denies fever, myalgias, nausea, vomiting, abdominal pain, chest pain, dysuria, hematuria, numbness and tingling, leg pain, difficulty walking, headache, visual disturbance, sore throat, rhinorrhea, and any other symptoms at this time."

This was because the span of the text containing the mention for "sore throat" did not include the word "denies" which negates the mention for "sore throat". This is an implementation issue which could be avoided by using sentence boundary detection,[45] and is something we are currently testing for the COVID-19 gazetteer. This issue led to mislabeling negative document-level mention of "sore throat" (cdc_sore_throat_n) as a positive document-level mention of "sore throat" (cdc_sore_throat_p).

Future work on the COVID-19 gazetteer includes expanding the experiments for COVID-19 gazetteer lexicon generation automation by increasing the pool of ED notes for COVID-19 patients. Lastly, the COVID-19 gazetteer is being ported across multiple compute nodes to improve runtime.

CONCLUSIONS

Compared to other annotation systems, the COVID-19 gazetteer demonstrates greater potential as a high-throughput annotation system for real-time processing of notes, therefore, providing an opportunity for clinicians to make more accurate time-sensitive decisions around patient care (eg, identifying symptoms of COVID-19 PUIs). With a continuously maintained and properly devised set of lexical rules, the COVID-19 gazetteer has the potential to perform similar to standard annotation systems for the task of symptom identification. Contrary to standard annotation systems the COVID-19 gazetteer has a considerably lower resource footprint and hence, could be deployed at medical sites lacking robust healthcare infrastructure. Thus, the COVID-19 gazetteer could be used as a fast, resource-efficient, and reliable tool for high-throughput real-time clinical decision support for COVID-19 or any other disease with well-defined symptomatology. It can be easily deployed in a large scale across a wide variety of healthcare settings for continuous surveillance of COVID-19 symptoms for prognostic purposes. In addition, it holds promise as a useful resource to study long-term sequelae of the disease in survivors (eg, PASC progression in COVID-19 survivors).

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONTRIBUTORS

HSS, MS—Project lead, lead developer of COVID-19 gazetteer, developer for UMLS NLP feature extraction, data analysis, data interpretation, writing, and critical revision. GMS, BS—Lead developer on Fairview COVID-19 pipeline process automation and NLP feature extraction, NLP-ADAPT-kube, study design, data analysis, data interpretation, writing, and critical revision. NEI, MD—Study design, architect on NLP feature extraction, ETL of extracted NLP features, data analysis, data interpretation, writing, and critical revision. MIL, MD—Study design, architect on NLP feature extraction, data interpretation, writing, and critical revision. MP, MD, MS—Study design, architect on NLP feature extraction, data analysis, data interpretation, writing, and critical revision. RLF, BS—Lead developer on NLP-ADAPT-kube and developer, writing, and critical revision. JS, PhD—Project advisor, study design, writing, and critical revision. RZ, PhD—Project advisor, study design, writing, and critical revision. BKK, BS—Lead developer on BioMedICUS and developer, writing, and critical revision. SL, PhD—Lead developer on MedTagger and developer, writing, and critical revision. HL, PhD—Project lead on MedTagger, writing, and critical revision. GBM, MD, PhD—Project advisor, study design, data interpretation, writing, and critical revision. CJT, MD—Project advisor, lead architect Fairview COVID-19 pipeline, architect of COVID-19 Patient Registry and NLP feature extraction, lead on study design, data analysis, data interpretation, writing, and critical revision. SP, PhD—Project advisor, NLP feature extraction conception, study design, data analysis, data interpretation, writing, and critical revision.

ACKNOWLEDGMENTS

We would like to extend our gratitude to Elizabeth Lindemann for her help in proofreading this manuscript. In addition, we would like to extend our gratitude to Eic Murray for providing us the infrastructure to help implement this project.

FUNDING

This research was supported by Fairview Health Services, the National Institutes of Health's National Center for Advancing Translational Sciences grant U01TR002062, the National Institutes of Health's National Heart, Lung, and Blood Institute's grant T32HL07741 (NEI), the Agency for Healthcare Research and Quality (AHRQ) R01HS026743 (MGU) and Patient-Centered Outcomes Research Institute (PCORI), grant K12HS026379 (S.S., C.J.T.). Additional support for MN-LHS scholars is offered by the University of Minnesota Office of Academic Clinical Affairs and the Division of Health Policy and Management, University of Minnesota School of Public Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of Fairview Health Services, the National Institutes of Health, AHRQ, PCORI, or Minnesota Learning Health System Mentored Career Development Program (MN-LHS). This grant was also supported by the University of Minnesota CTSA grant UL1TR000114.

Conflict of interest statement. None declared.

DATA AVAILABILITY

The data underlying this article cannot be shared publicly due to Health Insurance Portability and Accountability Act (HIPAA) rule.

REFERENCES

1. Chard K, Russell M, Lussier Y, *et al.* Scalability and cost of a cloud-based approach to medical NLP. In: 2011 24th International Symposium on Computer-Based Medical Systems (CBMS). 2011. 1–6. doi: 10.1109/CBMS.2011.5999166.
2. Miller TA, Avillach P, Mandl KD. Experiences implementing scalable, containerized, cloud-based NLP for extracting Biobank participant phenotypes at scale. *JAMIA Open* 2020; 3 (2): 185–9.
3. Liu H, Bielinski SJ, Sohn S, *et al.* An information extraction framework for cohort identification using electronic health records. *AMIA Jt Summits Transl Sci Proc* 2013; 2013: 149–53.
4. Wagholikar A, Zuccon G, Nguyen A, *et al.* Automated classification of limb fractures from free-text radiology reports using a clinician-informed gazetteer methodology. *Australas Med J* 2013; 6 (5): 301–7.
5. An introduction to named entity recognition in natural language processing - Part 1 and 2. Data community DC 2013. https://www.datacommunitydc.org/blog/2013/04/a-survey-of-stochastic-and-gazetteer-based-approaches-for-named-entity-recognition.
6. Elkin PL, Froehling D, Wahner-Roedler D, *et al.* NLP-based identification of pneumonia cases from free-text radiological reports. *AMIA Annu Symp Proc* 2008; 2008: 172–6.
7. Couto FM, Lamurias A. MER: a shell script and annotation server for minimal named entity recognition and linking. *J Cheminform* 2018; 10 (1): 58.
8. Meystre S, Haug PJ. Automation of a problem list using natural language processing. *BMC Med Inform Decis Mak* 2005; 5: 30.
9. Wen A, Fu S, Moon S, *et al.* Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. *NPJ Digit Med* 2019; 2: 130.
10. N3C. Rule Editor. 2021. https://ohnlp.github.io/ohnlptk/ie_editor.html. Accessed March 8, 2021.
11. Docker Inc. What is a container? A standardized unit of software. 2020. https://www.docker.com/resources/what-container. Accessed August 10, 2021.
12. Argo Project Authors. Argo Workflows & Pipeline. 2020. https://argoproj.github.io/workflows/. Accessed August 10, 2021.
13. Stephens KA, Au MA, Yetisgen M, *et al.* Leveraging UMLS-driven NLP to enhance identification of influenza predictors derived from electronic medical record data. *bioRxiv* 2020; 2020.04.24.058982. doi:10.1101/2020.04.24.058982.
14. CDC. Coronavirus Disease 2019 (COVID-19) – Symptoms. Cent. Dis. Control Prev. 2020. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html. Accessed February 15, 2021.
15. Coronavirus disease 2019 (COVID-19) - Symptoms and causes. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/coronavirus/symptoms-causes/syc-20479963. Accessed April 23, 2021.
16. He Y, Yu H, Ong E, *et al.* CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. *Sci Data* 2020; 7 (1): 181.
17. NLM. UMLS language system: statistics 2020AB release. 2020. https://www.nlm.nih.gov/pubs/techbull/nd20/nd20_umls_release.html. Accessed August 10, 2021.
18. Mikolov T, Chen K, Corrado G, *et al.* Efficient estimation of word representations in vector space. arXiv:13013781 [cs] [Published online ahead of print 6 September 2013]. http://arxiv.org/abs/1301.3781 Accessed August 10, 2021.
19. Pakhomov S, Finley GP, McEwan R, *et al.* Corpus domain effects on distributional semantic modeling of medical terms. *Bioinformatics* 2016; 32 (23): 3635–44.
20. Finzel R, Silverman G. *nlp-adapt-kube*. 2019. https://github.com/nlpie/nlp-adapt-kube. Accessed January 6, 2020.

21. Apache Foundation. UIMA Project. UIMA Proj. 2013. https://uima.apache.org. Accessed February 8, 2020.
22. Knoll B. *biomedicus*. UMN NLP?IE. 2019. https://github.com/nlpie/biomedicus. Accessed January 6, 2020.
23. Savova GK, Masanz JJ, Ogren PV, *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010; 17 (5): 507–13.
24. The National Institutes of Health. MetaMap. 2019. https://metamap.nlm.nih.gov. Accessed January 6, 2020.
25. Soysal E, Wang J, Jiang M, *et al.* CLAMP a toolkit for efficiently building customized clinical natural language processing pipelines. *J Am Med Inform Assoc* 2018; 25 (3): 331–6.
26. OHNLP/MedTagger. Open health natural language processing. 2021. https://github.com/OHNLP/MedTagger. Accessed March 4, 2021.
27. OHNLP/covid19ruleset. Open health natural language processing. 2021. https://github.com/OHNLP/covid19ruleset. Accessed March 4, 2021.
28. Matcher spaCy API Documentation. Matcher. https://spacy.io/api/matcher. Accessed February 23, 2021.
29. ExplosionAI. EntityRuler spaCy API Documentation. 2021. https://spacy.io/api/entityruler. Accessed February 5, 2021.
30. English spaCy Models Documentation. English. https://spacy.io/models/en. Accessed February 23, 2021.
31. negspaCy spaCy Universe. negspaCy. https://spacy.io/universe/project/negspacy. Accessed February 5, 2021.
32. nlpie/covid_symptom_gazetteer. GitHub. https://github.com/nlpie/covid_symptom_gazetteer. Accessed February 23, 2021.
33. Gamakaranage CSK, Hettiarachchi D, Ediriweera D, *et al.* Symptomatology of Coronavirus Disease 2019 (COVID-19) - lessons from a meta-analysis across 13 countries. 2021. doi: 10.21203/rs.3.rs-39412/v1.
34. CDC. COVID-19 and Your Health. Cent. Dis. Control Prev. 2020. https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects.html. Accessed March 8, 2021.
35. cTAKES 4.0 - Apache cTAKES - Apache Software Foundation. https://cwiki.apache.org/confluence/display/CTAKES/cTAKES+4.0. Accessed February 22, 2021.
36. Demner-Fushman D, Rogers WJ, Aronson AR. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. *J Am Med Inform Assoc* 2017; 24 (4): 841–4.
37. Pradhan S, Elhadad N, Chapman W, *et al.* SemEval-2014 task 7: analysis of clinical text. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: Association for Computational Linguistics; 2014: 54–62. doi:10.3115/v1/S14-2007.
38. Doğan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. *J Biomed Inform* 2014; 47: 1–10.
39. Vincze V, Szarvas G, Farkas R, *et al.* The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics* 2008; 9 Suppl 11: S9.
40. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011; 18 (5): 552–6.
41. Finzel RF, Silverman GM, Datar S, AIME 2020 - Tutorials and workshops: large scale ensembled NLP systems with docker and kubernetes. https://aime20.aimedicine.info/index.php/tutorials-and-workshops#tutorial2. Accessed February 22, 2021.
42. Misra D. Extracting patient narrative from clinical notes: implementing apache cTAKES at scale using apache spark. https://drive.google.com/drive/folders/1ngYeqkNWZNMLNpM69OFC9cDTEXYt5hPz. Accessed February 23, 2021.
43. Apache SparkTM - Unified Analytics Engine for Big Data. https://spark.apache.org/. Accessed February 23, 2021.
44. Devlin J, Chang M-W, Lee K, *et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In: NAACL-HLT. 2019.
45. Sentence boundary disambiguation. Wikipedia. 2020. https://en.wikipedia.org/w/index.php?title=Sentence_boundary_disambiguation&oldid=990546596. Accessed June 2, 2021.