Journal of the American Medical Informatics Association, 29(10), 2022, 1797–1806
https://doi.org/10.1093/jamia/ocac127
Advance Access Publication Date: 3 August 2022
Research and Applications

Research and Applications

A scoping review of publicly available language tasks in clinical natural language processing

Yanjun Gao [1], Dmitriy Dligach [2], Leslie Christensen [3], Samuel Tesch [3], Ryan Laffin [3], Dongfang Xu [4], Timothy Miller [4], Ozlem Uzuner [5], Matthew M. Churpek [1], and Majid Afshar [1]

[1]ICU Data Science Lab, Department of Medicine, School of Medicine and Public Health, University of Wisconsin, Madison, Wisconsin, USA, [2]Department of Computer Science, Loyola University Chicago, Chicago, Illinois, USA, [3]School of Medicine and Public Health, University of Wisconsin, Madison, Wisconsin, USA, [4]Computational Health Informatics Program, Boston Children's Hospital, Harvard University, Boston, Massachusetts, USA, and [5]Department of Information Sciences and Technology, George Mason University, Fairfax, Virginia, USA

Corresponding Author: Yanjun Gao, PhD, ICU Data Science Lab, School of Medicine and Public Health, University of Wisconsin, Madison, WI, USA; ygao@medicine.wisc.edu

Received 1 December 2021; Revised 16 June 2022; Editorial Decision 2 July 2022; Accepted 1 August 2022

ABSTRACT

Objective: To provide a scoping review of papers on clinical natural language processing (NLP) shared tasks that use publicly available electronic health record data from a cohort of patients.
Materials and Methods: We searched 6 databases, including biomedical research and computer science literature databases. A round of title/abstract screening and full-text screening were conducted by 2 reviewers. Our method followed the PRISMA-ScR guidelines.
Results: A total of 35 papers with 48 clinical NLP tasks met inclusion criteria between 2007 and 2021. We categorized the tasks by the type of NLP problems, including named entity recognition, summarization, and other NLP tasks. Some tasks were introduced as potential clinical decision support applications, such as substance abuse detection, and phenotyping. We summarized the tasks by publication venue and dataset type.
Discussion: The breadth of clinical NLP tasks continues to grow as the field of NLP evolves with advancements in language systems. However, gaps exist with divergent interests between the general domain NLP community and the clinical informatics community for task motivation and design, and in generalizability of the data sources. We also identified issues in data preparation.
Conclusion: The existing clinical NLP tasks cover a wide range of topics and the field is expected to grow and attract more attention from both general domain NLP and clinical informatics community. We encourage future work to incorporate multidisciplinary collaboration, reporting transparency, and standardization in data preparation. We provide a listing of all the shared task papers and datasets from this review in a GitLab repository.

Key words: natural language processing, clinical informatics, electronic health records, systematic review, clinical decision support

INTRODUCTION

Since the inception of the first Integrating Biology and the Bedside (i2b2) shared task in 2006, currently known as the National Natural Language Processing (NLP) Clinical Challenge (n2c2), the field of clinical NLP has advanced in clinical applications that rely on text from the electronic health record (EHR). Tasks with publicly avail-

© The Author(s) 2022. Published by Oxford University Press on behalf of the American Medical Informatics Association.
All rights reserved. For permissions, please email: journals.permissions@oup.com

able data (eg, shared tasks) provide a new avenue for advancing the state-of-the-art using publicly available datasets in a sector that is otherwise heavily regulated and protected from sharing patient data. In an editorial approximately a decade ago, Chapman et al[1] identified the major barriers to clinical NLP developments where shared tasks may provide a solution. Some of the challenges were a lack of data resources, including annotation tools, benchmarking and standardized metrics, reproducibility, and collaboration between the general NLP communities and health research communities.

Over the past decade, strides have been made with an increasing number and heterogeneity in clinical NLP tasks, and many organizers are leveraging publicly available EHR notes like the Medical Information Mart for Intensive Care (MIMIC).[2] MIMIC along with clinical notes from other health systems has overcome privacy and regulatory hurdles to enable the growth of language tasks to address important clinical problems with NLP solutions. The benefits of publicly available language tasks have become apparent with an opportunity for both clinical informatics (CI) and general domain NLP communities to tackle problems together and develop systems that may translate into applied tools in health systems. The body of language tasks continues to enable the growth with complex information extraction tasks ranging from early diagnoses (eg, substance abuse detection, phenotyping[3–5]) to clinical language understanding (eg, natural language inference[6,7]).

However, several challenges remain as transparency in the methods, clinical motivation, and standardization across annotation techniques and sample size determination remain highly variable. Our objective is to review papers describing clinical NLP shared tasks that use publicly available EHR data from a cohort of patients. We aim to examine the progress over the years and describe both barriers that we have overcome as well as challenges that remain in advancing clinical NLP. This scoping review will serve as a resource, accompanied by a GitLab repository (https://git.doit.wisc.edu/YGAO/public-available-clinical-nlp-tasks/-/tree/main/), for organizers and participants in the clinical NLP domain to quickly retrieve details on publicly available clinical tasks as well as identify gaps and opportunities for future tasks.

MATERIALS AND METHODS

The methods to conduct this scoping review adhered to standards described in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines for Scoping Reviews (PRISMA-ScR).[8] The search design identified new language tasks for clinical NLP using publicly available EHRs from a cohort of patients. The task required expert annotations to build a labeled corpus of data. The comparator is a test dataset with an evaluation metric for the task. The goal of the task was to provide a benchmark and enable the development of state-of-the-art (STOA) models for the task.

Literature search

The librarian (LC) performed a full, systematic review of the literature between January 1985 and September 2021. The search combined controlled vocabulary and title/abstract terms related to the shared language tasks in clinical NLP, focusing on publicly available datasets. The search was developed in PubMed, tested against a set of exemplar articles, and then translated into the following databases: (1) Embase (Scopus); (2) The Association for Computing Machinery (ACM) Guide to Computing Literature (ACM Digital Library); (3) Science Citation Index Expanded (Web of Science); (4) Conference Proceedings Citation Index-Science (Web of Science); and (5) Emerging Sources Citation Index (Web of Science). The metadata from the Association for Computational Linguistics (ACL) Anthology was downloaded separately and searched based on the database search strategies. The search strategies were peer-reviewed by 2 University of Wisconsin (UW)-Madison Science and Engineering librarians. All searches were performed on September 8, 2021 except for ACL, which was on September 1, 2021. No publication type, language, or date filters were applied. Results were downloaded to a citation management software (EndNote x9, Clarivate Analytics, Philadelphia, PA, USA) and underwent manual deduplication by the librarian. Unique records were uploaded to Rayyan screening platform[9] for independent review. The full query with search terms and Boolean operations for each database is detailed in Supplementary Appendix C.

Study selection

Study inclusion criteria were the following: (1) publicly available dataset for the shared task; (2) clinical NLP task; (3) novel benchmark metric; (4) models that were built and tested to establish state-of-the-art results for the novel benchmark metric; and (5) English-language research articles and tasks. Articles were excluded if the tasks were focused on the biomedical domain (genomics data, nonpatient data, data from clinical research databases including PubMed articles), subject-matter specific tasks without publicly available data, preprints or nonpeer-reviewed, and individual use-case systems not designed as a shared task. Multiple papers shared a data challenge with multiple tracks. For example, the 2014 i2b2/UTHealth shared task had 2 tracks, protected health information (PHI) deidentification and temporal identification of risk factors for heart disease. We analyzed each track as its own task, and some tasks consisted of multiple subtasks. If the subtask focused on a distinct clinical problem, we also considered each subtask its own task. We excluded subtasks when the data were not clinical text, and it was not related to clinical NLP.

Researchers with expertise in NLP and CI (YG and MA) performed a review of titles and abstracts for inclusion into full-text article review. The first 400 titles/abstracts were reviewed by the 2 reviewers in a blinded fashion and the Cohen's Kappa score for interannotator agreement was 0.83. The subsequent papers were divided and reviewed by each reviewer independently. Any disagreements or indeterminate decisions were resolved through discussion and consensus.

Data synthesis and summarization

Among the papers included in the scoping review, characteristics of the shared tasks were described and the data corpus metrics were summarized into tables. The following characteristics were provided: (1) publication date and location; (2) the type of NLP task and data source; (3) level of annotation; (4) participant details; (5) data corpus details; (6) number of citations for the task; and (7) evaluation metrics. Depending on where the task was published, we categorized each task as originating from the CI or general domain NLP community. Most papers explicitly defined the type of NLP problems the task addresses. For these papers, we followed their task definitions. For the remaining papers, we categorized the types based on the type of input and output, following the conventional definition in the general NLP domain. Therefore, we acknowledge that named entity recognition (NER) is a type of information extraction (IE), but we separated them because of how it appeared in the shared task paper. A systematic review protocol for our study was

submitted a priori to the PROSPERO database for prospectively registered systematic reviews. Our protocol was deemed as a scoping review because of the heterogeneity in tasks and evaluation metrics; therefore, protocol registration was not required, and we followed the guideline and checklist from the 2018 PRISMA-ScR (Supplementary Appendix B).[8]

RESULTS

Search results

Our search results identified 4489 abstracts for review after deduplication. After the first review phase with title/abstract screening, 99 papers met inclusion criteria for full-text review. During the full-text review phase, 68 papers were excluded and the most common reason for exclusion was not having publicly available data ($n = 24$). During the full-text review, another 5 papers were identified that were not part of the original query results. Thirty-five papers spanning 48 clinical NLP tasks between 2007 and 2021 were ultimately included for analysis. Figure 1 illustrates the selection process and results. All of the included papers were published in peer-reviewed CI and general domain NLP journals and conference proceedings.

General characteristics of included papers

The majority of tasks appeared in CI journals including the *Journal of the American Medical Informatics Association* (JAMIA; $n = 11$),[10–20] the *Journal of Biomedical Informatics* (JBI; $n = 5$),[21–25] and the *Journal of Medical Internet Research* (JMIR; $n = 2$).[5,26] The remaining papers were distributed across other health/clinical informatics journals and conference proceedings, including Artificial Intelligence in Medicine,[3] *Journal of Biomedical Semantics*,[27] *Drug Safety*,[28] and American Medical Informatics Association Symposium (AMIA, $n = 2$)[29,30] and World Congress on Medical and Health Informatics (MEDINFO).[31] In the general domain NLP community, the major proceedings included the Association of Computational Linguistics (ACL, $n = 2$),[32,33] International Conference on Language Resources and Evaluation (LREC, $n = 2$),[4,34] and Empirical Methods in Natural Language Processing (EMNLP, $n = 2$),[7,35] International Conference of the Cross-Language Evaluation Forum for European Language (or known as Conferences and Labs of Evaluation Forum, CLEF, $n = 3$),[36–38] and International World Wide Web Conference (WWW).[39] Some tasks were published in workshops such as the International Workshop on Semantic Evaluation (SemEval, $n = 2$),[40,41] Biomedical Natural Language Processing Workshop (BioNLP, $n = 2$),[6,42] and Workshop on Natural Language Processing for Medical Conversations.[43] One paper was published in the *Journal of Language Resources and Evaluation*.[44]

Authorship in the papers published in the CI and general domain NLP community did occasionally overlap, but we showed results separately for the 2 communities for the following reasons: (1) the peer review process and target audiences between CI and NLP publications were considerably different; (2) the 2 communities did not share the same publication index catalog; and (3) we aimed to uncover the differing motivations and goals for building clinical NLP systems. Several differences existed in the types of tasks shared between these 2 communities. Figure 2 illustrates the type of tasks and counts between 2007 and 2021 from the 2 communities. Overall, 28 of the tasks were published by the CI community, and 20 tasks were published by the general domain NLP community. The earliest shared task was published in a CI journal in 2007 (i2b2 Protected

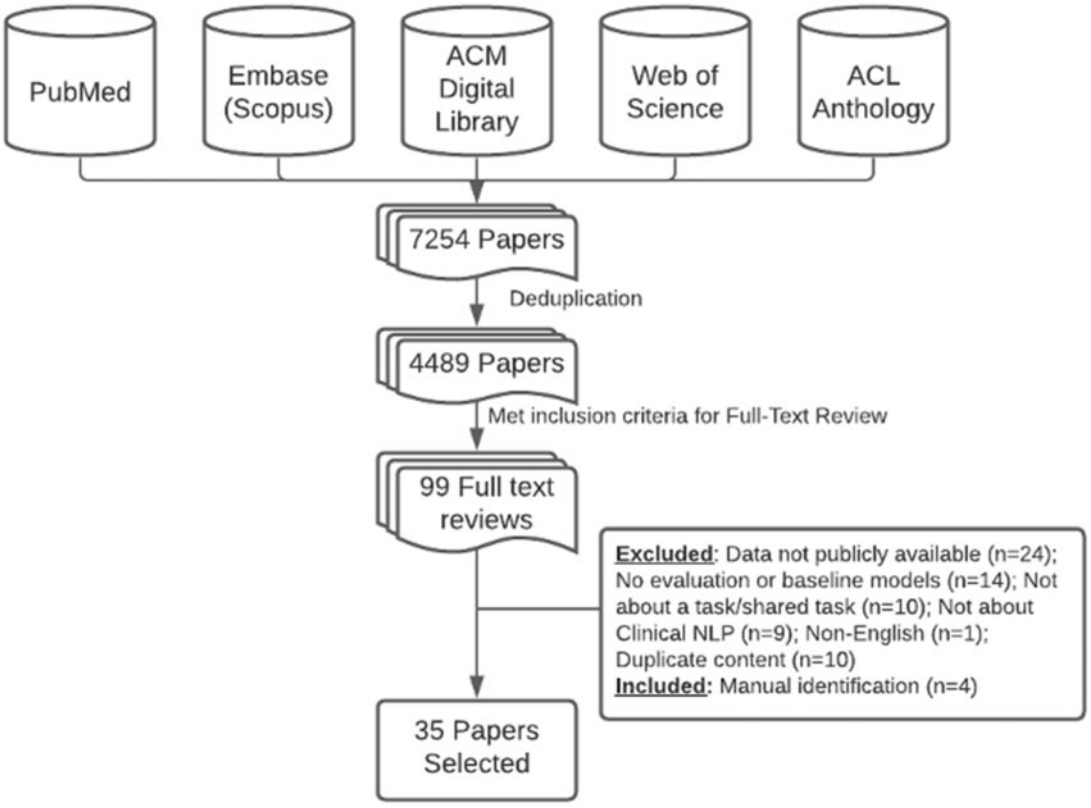

Figure 1. PRISMA diagram of our paper review process.

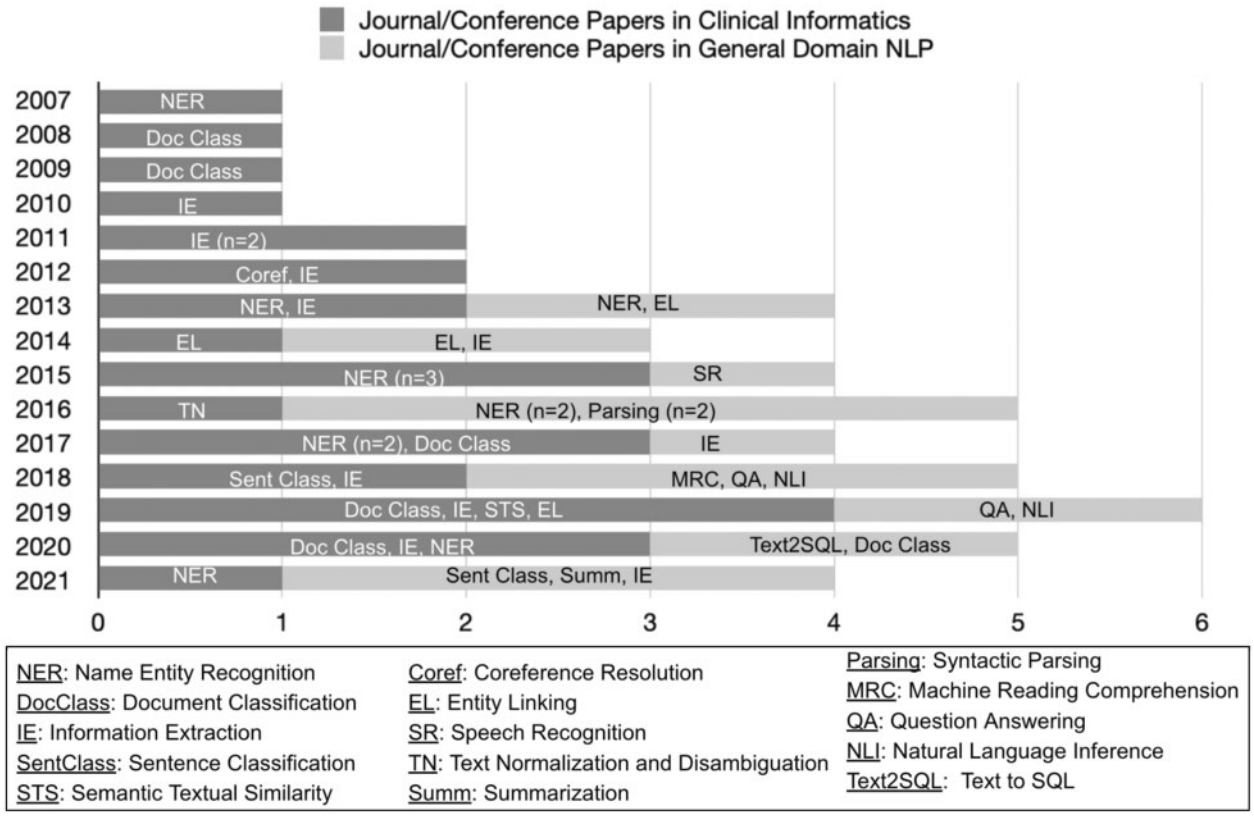

Figure 2. Types of tasks published in general domain NLP and clinical informatics journals and conference venues across years.

Health Information [PHI] De-Identification[10]) showing a longer history of developing clinical NLP tasks among the CI community. Six years later, the NLP community published its first clinical NLP task in the 2013 CLEF eHealth Task 2 Disorder Mention.[37] Interests from the general domain NLP community have increased over the years, representing the majority of tasks in 2021 (Summarization,[42] Action Item Extraction,[32] Assertion Detection[43]).

NER represented nearly a quarter of all tasks (25%, $n = 12$) with 18.75% ($n = 9$) and 6.25% ($n = 3$) in CI[3,5,10,16,17,21,22,24,28] and general domain NLP papers,[4,37,44] respectively. Other tasks that occurred frequently in CI were the broader IE tasks ($n = 11$),[13–16,19,28,31,36,41,43] Document Classification (DocClass; $n = 6$).[11,12,20,23,25,34] In the general domain NLP community, the types of tasks were distributed relatively evenly across Sentence Classification (SentClass; $n = 2$),[30,32] Entity Linking (EL; $n = 4$),[17,18,37,40] Syntactic Parsing (Parsing; $n = 2$),[4,44] Natural Language Inference (NLI; $n = 2$),[6,7] and Question Answering (QA).[6,35] Tasks that required text understanding and generation were proposed by general NLP community, such as Machine Reading Comprehension (MRC),[33] Summarization (Summ),[42] and Coreference Resolution (Coref).[15] A full description of the tasks and their definitions are detailed in Supplementary Appendix A.

Descriptions of included tasks and data

The characteristics of the tasks are shown in Table 1. We found that 38% of the NLP tasks were introduced with an intent for clinical decision making. Most of the clinical applications were NER tasks, introducing detection and identification of various medical conditions,[17] substance abuse,[3] medical risk factors,[21] medical events,[16,19] and PHI deidentification.[10,22,24] Phenotyping[44] introduced a corpus annotated with NER without identifying a specific clinical intent. Inconsistencies in defining NLP tasks occurred with 2 papers[4,5] that described phenotyping as an NER task and others[25,34] described it as a document classification task. IE was the second most frequent NLP task after NER (13.89%, $n = 9$), with some tasks focusing on time relation extraction,[14,16,41] and concept extraction,[13,14,19] others focusing on mention of substance uses,[28] and disorder.[36] Tasks without specific clinical applications were NLI,[6,7] MRC,[33] QA,[6,35] Summ,[42] Semantic Textual Similarity (STS),[26] Coref,15 Parsing,[4,44] Text2SQL,[39] and Speech Recognition (SR).[38] Most of these tasks were introduced by the general domain NLP community, except STS[26] and Coref.15

The data sources used to build the corpora were frequently derived from single health systems. Among them, the most frequent was from MIMIC,[2] which was from a large tertiary academic center in Boston and represented 31.91% ($n = 15$) of the tasks.[6,7,19,25,27,32,34,36,37,39,40,42,43] Other urban and academic health systems also contributed by releasing their data in a deidentified format including the following: Partners HealthCare (PHC, $n = 8$)[10–13,16,18,23,24,33]; Beth Israel Deaconess Medical Center (BIDMC, $n = 7$)[14,16,18,33]; University of Pittsburg Medical Center (UPMC, $n = 5$)[14,15,33]; University of Texas Health System (UTHealth, $n = 4$)[14,15,33]; Mayo Clinic (Mayo, $n = 3$)[5,26,41]; and University of Washington Harborview Medical Center (UW Harborview, $n = 3$).[4,25] All these data sources represented single centers that were tertiary academic medical centers.

Several papers were general in describing the note types as "EMR" or "EHR" without further specifying the type of note (eg, progress note, discharge summary, radiology report, etc.). For those papers, we denoted the type as "clinical notes" only ($n = 14$).[20–

Table 1. Overview of tasks and datasets, including the type of tasks, data source, and annotation units (*n*=number of tasks)

NLP task[a]	Publications[b]	Number of citations[c]	Impact on clinical decision making	Data source	Note type	Annotation unit
Entity linking	Pradhan et al[17]	128	General nonspecific	MIMIC	Clinical notes	Word token
	Henry et al[18]	14	General nonspecific	BIDMC, PHC	Discharge summaries	
	Suominen et al[37]	270	General nonspecific	MIMIC	Discharge summaries	
	Pradhan et al[40]	169	General nonspecific	MIMIC	Clinical notes	
Natural language inference	Abacha et al[6]	76	General nonspecific	MIMIC	History and physical admission	Sentence pairs
	Romanov et al[7]	113	General nonspecific	MIMIC	History and physical admission	
Text disambiguation and normalization	Mowery et al[27]	29	General nonspecific	MIMIC	Discharge summaries, electrocardiogram, echocardiogram, radiology report	Word token
Machine reading comprehension	Yue et al[33]	16	General nonspecific	PHC, UPMC, UTHealth, BIDMC	Clinical notes	Document
Question answering	Abacha et al[6]	76	General nonspecific	MIMIC	History and physical admission	Sentence pairs
	Pampari et al[35]	92	General nonspecific	PHC, BIDMC, MIMIC	Clinical notes	
Summarization	Abacha et al[42]	17	General nonspecific	MIMIC	Radiology reports	Document
Named entity recognition	Yetisgen et al[3]	11	Substance abuse detection	MTSamples	History and physical admission	Word token
	Klassen et al[4]	3	Phenotyping	UW HarborView	Radiology report	
	Shen et al[5]	7	Phenotyping	Mayo	History and physical admission	
	Uzuner et al[10]	505	PHI deidentification	PHC	Discharge summaries	
	Sun et al[16]	422	General nonspecific	PHC	Discharge summaries	
	Pradhan et al[17]	128	Disorder detection	MIMIC	Clinical notes	
	Stubbs et al[21]	127	Risk factor identification	UT Health	Clinical notes	
	Stubbs et al[22]	185	PHI deidentification	UT Health	Clinical notes	
	Stubbs et al[24]	86	PHI deidentification	PHC	Clinical notes	
	Jagannatha et al[28]	85	General nonspecific	UMass Memorial Medical Center	Clinical notes	
	Suominen et al[37]	270	Disorder detection	MIMIC	Discharge summaries	
	Savkov et al[44]	26	General nonspecific	UK National Health	Clinical notes	
Information extraction	Uzuner et al[13]	472	General nonspecific (concept)	PHC	Discharge summaries and progress notes	
	Uzuner et al[14]	1038	General nonspecific (concept, relation, assertion)	PHC	Discharge summaries	Word token
	Uzuner et al[15]	172	General nonspecific	PHC, BIDMC, UPMC	Discharge summaries	
	Sun et al[16]	422	General nonspecific (time)	PHC, BIDMC	Discharge summaries	
	Henry et al[19]	90	General nonspecific (concept)	MIMIC	Discharge summaries	
	Viani et al[31]	8	General nonspecific (time)	UK National Health	Clinical notes	
	Kelly et al[36]	146	Disorder mention	MIMIC	Discharge summaries, radiology report, electrocardiogram	
	Bethard et al[41]	74	General nonspecific (time)	Mayo	Clinical notes	
	van Aken et al[43]	1	General nonspecific (assertion)	MIMIC	Discharge summaries	

(continued)

Table 1. continued

NLP task[a]	Publications[b]	Number of citations[c]	Impact on clinical decision making	Data source	Note type	Annotation unit
Semantic textual similarity	Wang et al[26]	16	General nonspecific	Mayo	Clinical notes	Sentence pairs
Coreference resolution	Uzuner et al[15]	172	General nonspecific	PHC, BIDMC, UPMC	Discharge summaries	Word token
Syntactic parsing	Klassen et al[4]	3	General nonspecific	UW Harborview	Radiology report	Sentence
	Savkov et al[44]	26	General nonspecific	UK National Health	Clinical notes	
Sentence classification	Peng et al[30]	100	General nonspecific (negation)	Source unspecified	Radiology reports	Sentence
	Mullenbach et al[32]	3	Action item extraction[32]	MIMIC	Discharge summaries	
Document classification	Uzuner et al[11]	389	Smoking status classification	PHC	Discharge summaries	Document
	Uzuner et al[12]	266	Obesity classification	PHC	Discharge summaries	
	Stubbs et al[20]	36	Cohort selection for clinical trial	UTHealth	Clinical notes	
	Filannino et al[23]	37	Symptom severity prediction	PHC	Psychiatric evaluation records	
	Lybargr et al[25]	21	Phenotyping	MIMIC, UW HarborView	Discharge summaries and clinical notes	
	Moseley et al[34]	2	Phenotyping	MIMIC	Discharge summaries and nursing progress notes	
Others (Text2SQL, speech recognition)	Goeuriot et al[38]	85	General nonspecific	NCTA	Nursing handover data	Various
	Wang et al[39]	21	General nonspecific	MIMIC	Structured data in EHR	

[a]Lexical normalization refers to transforming text into a single canonical form that is different from the original word. Machine reading comprehension refers to answering a query given context information. Semantic textual similarity refers to generating semantic representation for pairs of sentences and measuring cosine similarity. Coreference resolution refers to identifying and mapping referring expressions to entities. Syntactic parsing refers to chunking, tokenizing, and labeling constituents in the text based on syntactic rules. See Supplementary Appendix for full description of all tasks. Text2SQL refers to converting text into SQL language. The note type "clinical note" was used for papers that only provided a general description of the note type as "EMR" or "EHR" without further specifying the type of note.

[b]We include the first author, title of the paper, and publication year.

[c]The citation count was collected through Google Scholar, on May 19, 2022.

MIMIC: Medical Information Mart for Intensive Care; VA: Veterans Affairs; PHC: Partners HealthCare; UMass: University of Massachusetts; BIDMC: Beth Israel Deaconess Medical Center; UPMC: University of Pittsburgh Medical Center; UK: United Kingdom; UW: University of Washington; PHI: Protected Health Information; EHR: Electronic Health Record; NCTA: Nursing Care Team Assistant; MTSamples: https://www.mtsamples.com/.

We use "clinical notes" to mark the data type when the paper did not specify the note type.

The full list of all papers and their shared tasks listed by year may be viewed and accessed at https://git.doit.wisc.edu/YGAO/public-available-clinical-nlp-tasks.

[22,24,26,28,31,33,40,41,44] Other papers gave clear specifications and discharge summary was the most frequent note type (35.41%. $n = 17$),[10–16,18,19,25,27,32,34,36,37,43] followed by radiology reports (14.58%, $n = 7$).[27,30,36,42] Other note types included history and physical admission, daily progress notes, electrocardiogram, echocardiogram, pathology reports, and psychiatric evaluation records.[5,7,23,27,34,36] Two tasks had note types that were different from all other tasks: Text2SQL included structured data with a goal of converting the tabular data into SQL language[39]; SR included audio records from nursing handover sessions with the intent of developing written text from spoken language.[38]

We found that more than a half of the tasks used data annotated at the token level (56.25%, $n = 27$).[3–5,10,13–19,21,22,24,27,28,37,40,44]For these tasks, tokens served as the basis for assigning labels. Tasks like NLI,[6,7] STS,[26] Parsing,[4,44] and SentClass[30,32] had annotations at the sentence level. Document-level annotation occurred for tasks in DocClass,[11,12,20,23,25,34] MRC,[33] QA,[35] and Summ.[42]

Descriptions of task participation, data size, and evaluation

We report details on participation, data split, and evaluation in Table 2. Among all tasks where participation information was available (when the task was hosted as a shared task), the number of participants ranged from 5 to 35 teams. Summ was only hosted once but published in 2021 as the task with the greatest number of teams ($n = 35$) submitting their systems.[42] The other 2 tasks with the most participants were the following: STS, published in 2020 and attracted 33 teams[26]; and DocClass with an average of 29.33 teams across 4 shared tasks.[11,12,20,23]

Sample sizes across the labels were highly variable and ranged from a few hundred manually annotated labels to semiautomated methods that produced several-fold more labels. None of the papers justified their sample size and they were simply reported as convenience samples. Further, not all tasks reported their data splits in the papers.[4,16,30] The units of dataset size were also heterogeneous, and sometimes not consistent with the annotation. For instance, annotation for EL tasks were created at lexical level, yet few papers reported the size regarding number of words and tags.[18] The task with the biggest corpora was QA, using the emrQA dataset.[35] Their annotation was generated semiautomatically on all i2b2 data. NLI also had large sets of sentence pairs, ranging between 11 000 and 14 000 for the train set, and 405 to 1400 for the test set.[6,7]

Accuracy and F1 were the 2 most frequent evaluation metrics. These 2 metrics focused on evaluating if predicted labels were correct against a gold standard, such as EL and NER. Most tasks in DocClass applied the F1 score with the exception of Mean Absolute Error reported in Ref.23 Some metrics were used for a specific NLP task, such as ROUGE[45] and BERTScore[46] for summarization evaluation[42]; Pearson Correlation for STS task.[26] Tasks in parsing used F1 as well as Unlabeled Attachment Score (UAS) and Labeled Attachment Score (LAS), 2 standard metrics evaluating predicted parsed labels.[4,44]

DISCUSSION

Lessons learned from past community interests and efforts

Our scoping review identified a total of 35 papers spanning multiple NLP tasks across both CI and general domain NLP communities. Among the oldest and most frequent tasks across both communities were NER, a token level task. A shift from token-level tasks (NER, EL, etc.) to document-level tasks (MRC, QA, Summ, etc.) was observed across the years with a growing interest for language understanding and text generation problems. As the NLP field continues to evolve since the introduction of transformers[49] and the capacity

Table 2. Overview of tasks, average number of participants across years, years range for publications, and evaluation metrics (*n*=number of tasks)

NLP task	Avg. number of participants	Publication years	Data split range		Evaluation metric
			Training	Test	
Entity linking	25.33 ($n = 4$)	2014–2020	50–199 notes	50–133 notes	Acc.
Natural language inference	17 ($n = 1$)	2018–2019	11k–14k pairs	405–1.4k pairs	Acc.
Text disambiguation and normalization	5 ($n = 1$)	2016	199 notes	99 notes	Acc.
Machine reading comprehension	NA	2020	91k queries	9.9k queries	Exact match, F1
Question answering	NA	2018–2021	658k–1M pairs	188k–296k pairs	Acc.
Summarization	35 ($n = 1$)	2021	91k notes	600 notes	ROUGE,[45] HOLMS,[46] BERTScore,[47] CheXBert[48]
Named entity recognition	15.57 ($n = 7$)	2007–2021	99–3.1k notes	117–896 notes	F1, Acc.
Information extraction	17.43 ($n = 7$)	2011–2019	300–876 notes	100–574 notes	Acc., F1
Semantic textual similarity	33 ($n = 1$)	2020	1.6k pairs	412 pairs	Pearson correlation
Coreference resolution	20 ($n = 1$)	2012	590 notes	388 notes	F1
Syntactic parsing	NA	2016	NA	NA	F1, UAS, LAS
Sentence classification	NA	2011–2021	518 notes	100 notes	F1
Document classification	29.33 ($n = 3$)	2008–2020	202–11k notes	86–8k notes	F1, MAE
Others (Text2SQL, speech recognition)	11.67 ($n = 2$)	2013–2020	Text2SQL: 37k records SR: 100 cases	Text2SQL: 4k records SR: 100 cases	Error rate percentage, Acc.

NA: Statistics Not Available; Acc.: Accuracy; F1: F-measure; UAS: Unlabeled Attachment Score; LAS: Labeled Attachment Score; MAE: Mean Absolute Error; SR: Speech Recognition.

to build large pretrained neural language systems increases,[50,51] the breadth of tasks is expected to grow. In recent years, the general domain NLP field has contributed natural language understanding and generation tasks,[50–54] but the CI domain remains primarily focused on NER[3,5,24] and DocClass.[20,23] This may represent a divergence in focus between the general domain community but is also due to the availability of data, which are larger and more accessible in the general domain.

The first publicly available task appeared in a CI journal by clinical NLP experts working at health systems affiliated with academia.[10] Several reasons for the earlier appearance by the CI community may include the difficulties in extracting clinical notes and privacy laws protecting patient data for sharing, which requires individuals with direct access to the EHR. The CI community of NLP experts brings together similar computational linguistic knowledge but they collaborate with healthcare providers to tackle the linguistic challenges in EHR data with a better understanding of the medical terms and clinical problems. This is also reflected in the longer history of shared tasks with a focus into a clinical problem (eg, deidentification,[10,22,24] clinical trial recruitment,[20] etc.). In general domain NLP, tasks organizers focused more on fundamental tasks like Parsing,[4,44] and NLI.[6,7] The general domain NLP tasks were typically authored by nonclinical experts with different motivations to develop new technologies in clinical text understanding and less attention to the clinical needs of health systems. To further advance the field of clinical NLP, both communities may benefit in designing tasks that are needed to advance the science in NLP but also delineate how it may be applied in clinical practice, such as a generation task to build a medical scribe that can reduce documentation burden for the provider.

Another major gap we identified is the limitations in generalizability of the data sources. The data are relatively homogeneous, deriving from mainly large, urban tertiary academic centers and mainly from single centers with a biased representation of the US population. Further, the notes derived from academic centers also contain a large proportion of notes in the EHR written by trainees, and may not be representative of community hospitals and health systems without trainees. The current environment with HIPAA privacy laws and resources for an Enterprise Data Warehouse largely limit the availability of data to centers. Currently, centers with informatics and computational expertise and resources support the data needed for public tasks and these remain limited to well-resourced academic centers. Moving forward, a concerted effort should be placed into sourcing data across the larger community of nonacademic centers from rural and urban demographics with a larger case mix and multicenter representation.

Selection of notes and benchmarks

Time-series data such as daily progress notes are rarely investigated in existing tasks. The discharge summary is the most frequent note type and typically the most detailed about hospital events and final diagnoses and treatments provided. While these may be useful for accomplishing certain NLP tasks, their clinical application in real-time remains limited. Augmented intelligence via clinical decision support systems frequently ingest data as events happen or use note types with time-sensitive appearance or repeated measures. Discharge summaries are typically the last documentation to resolve what happened during a hospital stay and may not be useful for augmented decision making. Other note types such as radiology and emergency department notes that are time-sensitive or daily progress notes that track disease and treatment plans each day are potentially more useful for real-time NLP applications, which is a goal for many researchers in the field.

F1 scores and overall accuracy are the most frequently used evaluation metrics, but they are only one component in reliability and validity testing. The extent to which a system measures what it is intended to measure requires multiple validity metrics. Criterion validity metrics with accuracy and correlation scores against reference standards are the de facto standard in tasks. However, construct and content validity are also important. Construct validity is needed when no universally accepted criterion exists to support the concept (or construct) being measured. This may require human evaluation to provide more than just frequentist statistics and better report benchmarks for natural language understanding and generation tasks. Content validity (or face validity), the extent to which the system predicted key words represents the gold standard concepts require more sophisticated approaches that can evaluate semantics and word order like the BERTScore.[47] In the clinical domain, meeting all the validity metrics may not be enough. Pragmatic testing through clinical applications with practice simulations that examine the system's effectiveness should also be considered in future tasks.

Issues in task/data preparation

Introducing, preparing, and releasing data for a new task requires complex thoughts and actions, yet details on data preparation are often neglected. In this review, we identified some issues above that may help to improve future task presentation. Additionally, a small number of papers presented results from pretrained models without explaining the training set which hinders reproducibility. We also found that the data split sizes reported for most papers did not match with the annotation units. Finally, none of the papers reported how they determined the minimum size of annotations needed to adequately train a model. Recall that even within the same type of tasks, the data size could range substantially from hundreds to thousands (eg, DocClass). Although it is widely known that annotations are limited to the resources (time, budget, etc.), not knowing the minimum sample size raises a crucial question about result reliability: will the model performance trained on this dataset be trust-worthy? Models developed for tasks like NLI, MRC, and QA are data hungry and the minimum sample size should be determined a priori, as these tasks require deeper understanding in semantics and relations. We believe by addressing these issues, researchers could make more robust contributions to clinical NLP.

Several limitations occurred in our study. First, our literature search may have missed shared tasks that were in preprint and awaiting acceptance into peer-review. We hope to share more recent tasks as they become available in our GitLab repository. Second, we focused only on English language tasks but the NLP community may be further along in certain tasks for other languages. Lastly, our focus is on the original publication that proposed a task with a STOA model; therefore, the work that improves the STOA performance is not in the scope of this work.

CONCLUSION

The interest in introducing and participating in clinical NLP tasks grows as more tasks surface each year. The breadth of tasks is also increasing with topics varying from tasks with specific clinical applications to those facilitating clinical language understanding and reasoning. Undoubtedly, the field will continue to grow and attract more researchers from both general NLP domain and the CI com-

munity. We encourage future work on proposing shared tasks to overcome barriers in community collaboration, reporting transparency, and consistency of data preparation. As a resource to the community, we provide a listing of all the publicly available tasks from this review in our GitHub repository at (https://git.doit.wisc.edu/YGAO/public-available-clinical-nlp-tasks).

FUNDING

This work was supported by NIH/NIDA grant number R01DA051464 (to MA), NIH/NIGM grant number R01 HL157262 (to MMC), NIH/NLM grant numbers R01LM012973 and R01LM012918 (to TIM), NIH NLM grant number R01LM010090 (to TM and DD), and NIH/NLM grant number R13LM013127 (to OU).

AUTHOR CONTRIBUTIONS

YG, MA, DD, MMC, and TM contributed to the design of the research. MA, LC, ST, and RL contributed to the data collection. YG and MA processed the data, performed analyses on the results and drafted the manuscript with input from DD, MMC, TM, OU, and DX. All authors discussed the results and contributed to the final manuscript.

SUPPLEMENTARY MATERIAL

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

ACKNOWLEDGMENTS

We thank David Bloom and Anne Glorioso for their reviews of the search query as computer science librarians at the University of Wisconsin—Madison.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

A listing of all the publicly available tasks from this review will be available on a GitLab repository https://git.doit.wisc.edu/YGAO/public-available-clinical-nlp-tasks.

REFERENCES

1. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc* 2011; 18 (5): 540–3.
2. Johnson AE, Pollard TJ, Shen L, *et al*. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; 3 (1): 1–9.
3. Yetisgen M, Vanderwende L. Automatic identification of substance abuse from social history in clinical text. In: Conference on Artificial Intelligence in Medicine in Europe. Springer; 2017: 171–81; Vienna, Austria.
4. Klassen P, Xia F, Yetisgen-Yildiz M. Annotating and detecting medical events in clinical notes. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16); 2016: 3417–21; Portorož, Slovenia.
5. Shen F, Liu S, Fu S, *et al*. Family history extraction from synthetic clinical narratives using natural language processing: overview and evaluation of a challenge data set and solutions for the 2019 National NLP Clinical Challenges (n2c2)/Open Health Natural Language Processing (OHNLP) competition. *JMIR Med Inform* 2021; 9 (1): e24008.
6. Abacha AB, Shivade C, Demner-Fushman D. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In: Proceedings of the 18th BioNLP Workshop and Shared Task; 2019: 370–9; Florence, Italy.
7. Romanov A, Shivade C. Lessons from natural language inference in the clinical domain. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018: 1586–96; Brussels, Belgium.
8. Tricco AC, Lillie E, Zarin W, *et al*. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med* 2018; 169 (7): 467–73.
9. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. *Syst Rev* 2016; 5 (1): 1–10.
10. Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc* 2007; 14 (5): 550–63.
11. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc* 2008; 15 (1): 14–24.
12. Uzuner O. Recognizing obesity and comorbidities in sparse data. *J Am Med Inform Assoc* 2009; 16 (4): 561–70.
13. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc* 2010; 17 (5): 514–8.
14. 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.
15. Uzuner O, Bodnari A, Shen S, Forbush T, Pestian J, South BR. Evaluating the state of the art in coreference resolution for electronic medical records. *J Am Med Inform Assoc* 2012; 19 (5): 786–91.
16. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *J Am Med Inform Assoc* 2013; 20 (5): 806–13.
17. Pradhan S, Elhadad N, South BR, *et al*. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *J Am Med Inform Assoc* 2015; 22 (1): 143–54.
18. Henry S, Wang Y, Shen F, Uzuner O. The 2019 national natural language processing (NLP) clinical challenges (n2c2)/Open health NLP (OHNLP) shared task on clinical concept normalization for clinical records. *J Am Med Inform Assoc* 2020; 27 (10): 1529–37.
19. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *J Am Med Inform Assoc* 2020; 27 (1): 3–12.
20. Stubbs A, Filannino M, Soysal E, Henry S, Uzuner O. Cohort selection for clinical trials: n2c2 2018 shared task Track 1. *J Am Med Inform Assoc* 2019; 26 (11): 1163–71.
21. Stubbs A, Kotfila C, Xu H, Uzuner O. Identifying risk factors for heart disease over time: overview of 2014 i2b2/UTHealth shared task Track 2. *J Biomed Inform* 2015; 58: S67–77.
22. Stubbs A, Uzuner O. Annotating longitudinal clinical narratives for de-identification: the 2014 i2b2/UTHealth corpus. *J Biomed Inform* 2015; 58: S20–9.
23. Filannino M, Stubbs A, Uzuner O. Symptom severity prediction from neuropsychiatric clinical records: overview of 2016 CEGS N-GRID shared tasks Track 2. *J Biomed Inform* 2017; 75: S62–70.
24. Stubbs A, Filannino M, Uzuner O. De-identification of psychiatric intake records: overview of 2016 CEGS N-GRID shared tasks Track 1. *J Biomed Inform* 2017; 75: S4–S18.
25. Lybarger K, Ostendorf M, Yetisgen M. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. *J Biomed Inform* 2021; 113: 103631.
26. Wang Y, Fu S, Shen F, Henry S, Uzuner O, Liu H. The 2019 n2c2/OHNLP track on clinical semantic textual similarity: overview. *JMIR Med Inform* 2020; 8 (11): e23375.

27. Mowery DL, South BR, Christensen L, *et al.* Normalizing acronyms and abbreviations to aid patient understanding of clinical texts: ShARe/CLEF eHealth Challenge 2013, Task 2. *J Biomed Semantics* 2016; 7: 43.
28. Jagannatha A, Liu F, Liu W, Yu H. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). *Drug Saf* 2019; 42 (1): 99–111.
29. Uzuner O. Second i2b2 workshop on natural language processing challenges for clinical records. In: AMIA. Annual Symposium Proceedings. AMIA Symposium; 2008: 1252–3.
30. Peng Y, Wang X, Lu L, Bagheri M, Summers R, Lu Z. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Jt Summits Transl Sci Proc* 2018: 188–96; Washington, DC.
31. Viani N, Kam J, Yin L, *et al.* Annotating temporal relations to determine the onset of psychosis symptoms. *Stud Health Technol Inform* 2019; 264: 418–22.
32. Mullenbach J, Pruksachatkun Y, Adler S, *et al.* CLIP: a dataset for extracting action items for physicians from hospital discharge notes. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); 2021: 1365–78; Online: Association for Computational Linguistics. Available from: https://aclanthology.org/2021.acl-long.109.
33. Yue X, Gutierrez BJ, Sun H. Clinical reading comprehension: a thorough analysis of the emrQA dataset. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2020: 4474–86; Online.
34. Moseley ET, Wu JT, Welt J, *et al.* A corpus for detecting high-context medical conditions in intensive care patient notes focusing on frequently readmitted patients. In: Proceedings of the 12th Language Resources and Evaluation Conference; 2020: 1362–7; Marseille, France: European Language Resources Association. Available from: https://aclanthology.org/2020.lrec-1.170.
35. Pampari A, Raghavan P, Liang J, Peng J. emrQA: a large corpus for question answering on electronic medical records. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018: 2357–68; Brussels, Belgium.
36. Kelly L, Goeuriot L, Suominen H, *et al.* Overview of the share/clef ehealth evaluation lab 2014. In: *International Conference of the Cross-Language Evaluation Forum for European Languages*; 2014: 172–91; Springer; Sheffield, United Kingdom.
37. Suominen H, Salantera S, Velupillai S, *et al.* Overview of the ShARe/CLEF eHealth evaluation lab 2013. In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer; 2013: 212–31; Valencia, Spain.
38. Goeuriot L, Kelly L, Suominen H, *et al.* Overview of the CLEF eHealth Evaluation Lab 2015. In: Mothe J, Savoy J, Kamps J, Pinel-Sauvagnat K, Jones G, San Juan E, Capellato L, Ferro N, editors. *Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2015*. Lecture Notes in Computer Science. Vol. 9283. Cham: Springer; 2015.
39. Wang P, Shi T, Reddy CD. Text-to-SQL generation for question answering on electronic medical records. 2020.
40. Pradhan S, Chapman W, Man S, Savova G. Semeval-2014 task 7: analysis of clinical text. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval); 2014; CiteSeer.
41. Bethard S, Savova G, Palmer M, *et al.* SemEval-2017 task 12. Clinical TempEval. 2017; Dublin, Ireland.
42. Abacha AB, M'Rabet Y, Zhang Y, Shivade C, Langlotz C, Demner-Fushman D. Overview of the MEDIQA 2021 shared task on summarization in the medical domain. In: Proceedings of the 20th Workshop on Biomedical Language Processing; 2021: 74–85; Online.
43. van Aken B, Trajanovska I, Siu A, Mayrdorfer M, Budde K, Loeser A. Assertion detection in clinical notes: medical language models to the rescue? In: Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations; 2021: 35–40; Online.
44. Savkov A, Carroll J, Koeling R, Cassell J. Annotating patient clinical records with syntactic chunks and named entities: the Harvey Corpus. *Lang Resour Eval* 2016; 50 (3): 523–48.
45. Lin CY. Rouge: a package for automatic evaluation of summaries. In: Text Summarization Branches Out; 2004: 74–81.
46. M'Rabet Y, Demner-Fushman D. HOLMS: alternative summary evaluation with large language models. In: Proceedings of the 28th International Conference on Computational Linguistics; 2020: 5679–5688; Barcelona, Spain (Online).
47. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. In: International Conference on Learning Representations; 2019; New Orleans, LA.
48. Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren M. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020: 1500–1519; Online.
49. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. In: Advances in Neural Information Processing Systems; 2017: 5998–6008; Long Beach, CA.
50. Devlin J, Chang MW, Lee K, Toutanova K. 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); 2019: 4171–86; Minneapolis, MN.
51. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019: 3615–20; Association for Computational Linguistics; Hong Kong, China.
52. Radford A, Narasimhan K. Improving language understanding by generative pre-training. 2018.
53. Roberts A, Raffel C, Shazeer N. How much knowledge can you pack into the parameters of a language model? In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020: 5418–26; Online.
54. Zhang J, Zhao Y, Saleh M, Liu P. Pegasus: pre-training with extracted gap-sentences for abstractive summarization. In: International Conference on Machine Learning; 2020: 11328–39; PMLR; Online.