Journal of the American Medical Informatics Association, 27(10), 2021, 1529–e1
doi: 10.1093/jamia/ocaa106
Advance Access Publication Date: 24 September 2020
Research and Applications

Research and Applications

The 2019 n2c2/UMass Lowell shared task on clinical concept normalization

Yen-Fu Luo[1,*], Sam Henry [2,*], Yanshan Wang[3], Feichen Shen[3], Ozlem Uzuner [2,4,5,+], and Anna Rumshisky[1,5,+]

[1]Department of Computer Science, University of Massachusetts Lowell, Lowell, Massachusetts, USA, [2]Department of Information Sciences and Technology, George Mason University, Fairfax, Virginia, USA, [3]Department of Health Sciences Research, Mayo Clinic, Rochester, New York, USA, [4]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA and [5]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

*Co-first authors.
+Co-last authors.
Corresponding Author: Sam Henry, Luter Hall 325, 1 Avenue of the Arts, Newport News, VA 23606, USA (samuel.henry@cnu.edu)

Received 10 February 2021; Revised 1 May 2021; Editorial Decision 12 May 2021; Accepted 14 May 2021

ABSTRACT

Objective: The n2c2/UMass Lowell spin-off shared task focused on medical concept normalization (MCN) in clinical records. This task aimed to assess state-of-the-art methods for matching salient medical concepts from clinical records to a controlled vocabulary. We describe the task and the dataset used, compare the participating systems, and identify the strengths and limitations of the current approaches and directions for future research.
Materials and Methods: Participating teams were asked to link preselected text spans in discharge summaries (henceforth referred to as concept mentions) to the corresponding concepts in the SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) and RxNorm vocabularies from the Unified Medical Language System. The shared task used the MCN corpus created by the organizers, which maps all mentions of problems, treatments, and tests in the 2010 i2b2/VA challenge data to the Unified Medical Language System concepts. Submitted systems represented 4 broad categories of approaches: cascading dictionary matching, cosine distance, deep learning, and retrieve-and-rank systems. Disambiguation modules were common across all approaches.
Results: A total of 33 teams participated in the shared task. The best-performing team achieved an accuracy of 0.8526. The median and mean performances among all teams were 0.7733 and 0.7426, respectively.
Conclusions: Overall performance among the top 10 teams was high. However, particularly challenging for all teams were mentions requiring disambiguation of misspelled words, acronyms, abbreviations, and mentions with more than 1 possible semantic type. Complex mentions of long, multiword terms were also challenging and, in the future, will require better methods for learning contextualized representations of concept mentions and better use of domain knowledge.

Key words: Natural language processing, clinical narratives, machine learning, concept normalization

INTRODUCTION

Secondary use of electronic health records for observational medical research has seen a dramatic rise in recent years, fueled by the improvement of predictive modeling techniques using machine learning (ML) methods.[1–5] Such retrospective research has had a transformative effect on a number of clinical applications, from disease phenotyping and mapping disease trajectories, to identifying high-risk patients and predictive modeling of patient outcomes, and informing practice.[6–10] Predictive models developed for such tasks often use a combination of structured and unstructured data, in which the latter includes narrative provider notes (discharge summaries, nursing notes, pathology reports, etc .).

It is widely acknowledged that clinical narrative from provider notes often contain information uniquely suited to improve predictive modeling for clinical research.[11–14] However, the use of such models that utilize clinical narrative features is often hampered by high variability of linguistic expressions for the same concept. For example, "myocardial infarction," "heart attack," and "MI" may refer to the same concept; indeed, in the Unified Medical Language System (UMLS),[15] they all map to the same concept unique identifier (CUI) *C0027051*. At the same time, ambiguity poses additional problems, as identical surface expressions are commonly used to refer to completely different concepts. For example, "transport" may be used to refer to a cell function (UMLS CUI *C0005528*) or to an activity (*C0206243*).

A common consequence of this is that models fail to generalize across different patient records, which ultimately prevents their deployment and integration into practice. Modern data-hungry natural language processing methods that use deep learning techniques to overcome this obstacle, learning generalizable representations from large quantities of text, such as the popular BERT model,[16] which learns contextualized term-level embedding, are often not applicable in the clinical domain, as clinical records are Health Insurance Portability and Accountability Act–regulated and cannot be easily shared. Recent attempts to leverage such methods by combining publicly available de-identified clinical narratives with other biomedical text[17,18] continue to suffer from insufficient access to the necessary quantities of clinical text.

The task of medical concept normalization (MCN) attempts to solve this problem more directly by linking or associating all mentions of medical concepts in clinical narrative to a standardized vocabulary of concepts using manually constructed in-domain knowledge sources. The goal of such normalization is to enable predictive models that rely on clinical text to generalize better across different patient records. MCN is also often a component of information retrieval (IR) and information extraction systems. Such systems perform named entity recognition to identify mentions of interest, followed by MCN to normalize the identified mentions.

This article describes the shared task challenge on normalization of medical concepts in clinical records organized by UMass Lowell as a community-led spin-off of the National NLP Clinical Challenges (n2c2). The challenge used the MCN corpus[19] created by the organizers using discharge summaries from the 2010 i2b2/VA clinical concept data,[20] in which all mentions of problems, treatments, and tests were normalized to the corresponding concepts (CUIs) in a subset of the UMLS comprising SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms)[21] and RxNorm.[22]

Until recently, ML approaches to medical concept normalization for clinical records have been limited, with rare exceptions,[23] and dictionary lookup almost universally selected as the best method.[24–26] Given that ML approaches have been pushing forward state-of-the-art performance on many clinical informatics tasks, one of the goals of the challenge was to determine whether effective ML approaches could be developed for this task that would improve upon dictionary lookup. Thirty-three teams from around the world participated in the challenge, using UMLS-normalized data from the MCN corpus to develop and evaluate automated systems.

Privacy is a concern in the clinical domain and often prevents data sharing. This makes the availability of annotated clinical corpora scarce[27] and makes direct comparison between MCN methodologies difficult. For that reason, while there have been a number of efforts to create annotated data for MCN in the biomedical and consumer health domains, normalization data for the clinical domain have remained scarce. Previous clinical MCN shared tasks have focused on a limited set of concepts and dealt only with the normalization of diseases and disorders.[28–30] The 2019 n2c2/UMass Lowell shared task addresses these problems by making the MCN corpus[19] publicly available, which allows a direct comparison of different approaches to MCN on clinical narrative from different institutions, using a data that cover a larger breadth of concepts than previous efforts.[19]

In this article, we first describe previous MCN efforts. Next, we describe the 2019 n2c2/UMass Lowell shared task track 3 dataset, the task, and each of the participating systems. Last, we present the results, perform an error analysis, and conclude by identifying directions for future research.

BACKGROUND AND SIGNIFICANCE

Table 1 summarizes the publicly available MCN datasets and shared tasks in the biomedical, consumer health, and clinical domains. The vocabulary to which mentions are normalized varies, and includes EntrezGene[42]; MeSH (Medical Subject Headings)[43]; MedDRA (Medical Dictionary for Regulatory Activities)[44,45] MEDIC (MErged DIsease voCabulary),[46] which combines MeSH and the OMIM (Online Mendelian Inheritance in Man)[47] vocabularies; SNOMED CT[21]; SIDER (SIDe Effect Resource) 4[48]; Australian Medicines Terminology[49]; and RxNorm.[22]

Within the biomedical domain, there have been several BioCreative challenges (https://biocreative.bioinformatics.udel.edu/) focused on MCN. BioCreative tasks I,[31] II,[32] and III[33] contained tracks focused on gene normalization. BioCreative V CDR task[34] focused on normalizing chemicals, diseases, and their interactions. Text Analysis Conference (TAC) 2017 (https://bionlp.nlm.nih.gov/tac2017adversereactions/)[35] presented a track focused on normalizing adverse drug reactions (ADRs) from drug labels. In addition to these, the National Center for Biotechnology Information disease corpus (https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/),[37] which extends the Arizona Disease Corpus,[36] focused on normalizing disorders in PubMed articles.

The focus within the consumer health domain has been on ADRs. Datasets for this purpose included SMM4H (Social Media Mining for Health) Shared Tasks (https://healthlanguageprocessing.org/sharedtask2/)[38]; the TwADR-S (https://zenodo.org/record/27354)[39] and TwADR-L (https://zenodo.org/record/55013)[40] datasets, which focus on (text snippets of) tweets; and the CSIRO Adverse Drug Event Corpus (Cadec) (https://data.csiro.au/dap/landingpage?pid=csiro%3A10948),[41] which studies a wider range of concept types including ADRs, diseases, drugs, symptoms, and findings in medical forum posts from Ask a Patient.

The biomedical and consumer health domains share some similarities with the clinical domain. However, clinical data have many unique characteristics that make annotated clinical data essential. Un-

Table 1. Summary of medical concept normalization datasets and shared tasks

Task	Domain	Data Source	Concepts	Vocabulary
BioCreative I Task 1B[31]	Biomedical	MEDLINE	Fly, mouse, and yeast genes	Organizer provided
BioCreative II Task 1B[32]	Biomedical	MEDLINE	Human genes	EntrezGene
BioCreative III GN[33]	Biomedical	PMC full text	Genes	EntrezGene
BioCreative V CDR Task A[34]	Biomedical	PubMed	Chemicals, diseases, chemical-disease interactions	MeSH
TAC 2017[35]	Biomedical	Drug labels	ADRs	MedDRA
AZDC Corpus[36]	Biomedical	PubMed	Disorders	UMLS
NCBI Corpus[37]	Biomedical	PubMed	Disorders	MEDIC
SMM4H 2017 Task 3[38]	Consumer health	Twitter	ADRs	MedDRA
TwADR-S[39]	Consumer health	Twitter	ADRs	SNOMED CT
TwADR-L[40]	Consumer health	Twitter	ADRs	SIDER 4
CADEC Corpus[42]	Consumer health	Online health forum	ADRs, diseases, drugs, symptoms, findings	SNOMED CT, MedDRA, AMT
ShARe/CLEF 2013 Tracks 1b and 2[28]	Clinical	Clinical records	Disorders	SNOMED CT
SemEval-2014 Task 7B[29]	Clinical	Clinical records	Disorders	SNOMED CT
SemEval-2015 Task 14 Tracks 1 and 2a[30]	Clinical	Clinical records	Disorders	SNOMED CT
2019 n2c2 Track 3	Clinical	Clinical records	Problems, treatments, tests	SNOMED CT, RxNorm

ADR: adverse drug reaction; AMT: Australian Medicines Terminology; AZDC: Arizona Disease Corpus; MedDRA: Medical Dictionary for Regulatory Activities; MeSH: Medical Subject Headings; NCBI: National Center for Biotechnology Information; SIDER: SIDe Effect Resource; SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.

fortunately, owing in large part to privacy concerns, there is a scarcity of publicly available clinical MCN corpora. To the best of our knowledge, only 3 shared tasks, ShARe/CLEF eHealth 2013 (https://sites.google.com/site/shareclefehealth/),[28] SemEval-2014 Task 7B (http://alt.qcri.org/semeval2014/task7/),[29] and SemEval-2015 Task 14 (http://alt.qcri.org/semeval2015/task14/),[30] have focused on clinical MCN and produced publicly available corpora. These tasks however, are limited. They focus on just a single concept type, *disorders*, and all use a corpus derived from the same institution. Specifically, they use the ShARe corpus (http://share.healthnlp.org), which originated from the MIMIC II (Medical Information Mart for Intensive Care II) database (http://mimic.physionet.org)[50] and consists of discharge summaries, electrocardiogram, echocardiogram, and radiology reports. While in some respects it is beneficial that these tasks build on each other by iteratively adding more annotations to a common dataset, limited institutional coverage of the resulting dataset constrains the generalizability of approaches developed on it.

In contrast to these tasks, 2019 n2c2/UMass Lowell shared task track 3 focused on normalizing a broad set of salient medical concepts, including medical problems, treatments, and tests. The inclusion of treatments in the annotation also means that an additional vocabulary, RxNorm[22] is used along with SNOMED CT in this challenge. At the same time, this shared task expands institutional coverage by using data from the 2010 i2b2/VA shared task dataset[20] which includes records from Partners HealthCare, in addition to the commonly used Beth Israel Deaconess Medical Center (via MIMIC II) data. Furthermore, the number of mentions without a mapping (CUI-less mentions) has been greatly reduced from around 30% in previous MCN challenges to 2.7% in this challenge. This reduction is due to the compositional annotation approach,[19] in which CUI-less mentions are split into multiple smaller spans. For example: the mention "Breast or ovarian cancer" may be annotated as 2 CUIs, [*C0006142, breast cancer*] and [*C0029925, ovarian cancer*], and the mention "Left breast biopsy" can be mapped to the CUIs [*C0222601, left breast*] and [*C0005558, biopsy*]. The result is a dataset that covers a broader set of concepts, expands institutional coverage, and minimizes CUI-less mentions. This allows a more effective assessment of MCN performance and a more comprehensive characterization of the state of the art for clinical MCN.

MATERIALS AND METHODS

Data

2019 n2c2/UMass Lowell shared task track 3 used the MCN corpus developed by Luo et al.[19] This corpus added normalization annotations to a subset of the 2010 i2b2/VA shared task dataset.[20] The 2010 i2b2/VA challenge identified mentions corresponding to problem, treatment, and test concepts found in narratives. MCN corpus mapped these mentions to CUIs from the controlled vocabularies, RxNorm[22] and SNOMED CT (SNOMEDCT_US)[21] within the 2017AB version of the UMLS.[15] SNOMED CT is a comprehensive vocabulary for clinical terminology and RxNorm is a comprehensive vocabulary focusing on clinical drugs and medications. Overall, the MCN corpus consists of 100 de-identified discharge summaries originating from Beth Israel Deaconess Medical Center and Partners Healthcare. These records contain 13 609 mentions, 3791 distinct CUIs, and 368 CUI-less mentions. Of the 3791 distinct CUIs, 244 are found in both SNOMED CT and RxNorm, while 3331 and 216 are found exclusively in SNOMED CT and RxNorm, respectively.

Shared task setup

For the 2019 n2c2/UMass Lowell shared task track 3, the MCN data were split into training and test sets. Each set contained 50 discharge summaries. The total number of mentions, CUIs, and number of CUI-less mentions were similar for both sets. The training set contained 6684 mentions, 2330 CUIs, and 151 CUI-less mentions. The test set contained 6925 mentions, 2578 CUIs, and 217 CUI-less mentions.

We released the training data, its gold standard annotations, and an evaluation script to participants after completion of a data use agreement via an online Web portal (https://n2c2.dbmi.hms.harvard.edu/track3). Teams were given an approximate 2-month development period before we released the test data. Gold standard

annotations on test data were withheld. Teams were given 2 days after the release of test data to submit their system's output via the online Web portal. We performed the final evaluation and posted the rankings of all teams online. We released the gold standard for the test data once the results were posted.

Annotation guidelines

Detailed annotation guidelines are reported by Luo et al.[19] Annotation consisted of double annotation by 4 pharmacy and nursing students followed by adjudication by a certificated professional medical coder. Annotators used the MAE annotation tool.[51] They were instructed to map every mention to a single CUI. Multiple CUIs per span were not allowed. As mentioned previously, CUI-less mentions were tackled by a compositional annotation approach,[19] which split these mentions into multiple smaller spans. Where synonymous CUIs existed for a span, the CUI was standardized across the corpus.

Preadjudication interannotator agreement (IAA) was calculated as the accuracy of the annotators over all annotated mentions. Postadjudication IAA is similar to preadjudication IAA but adds synonymous CUIs to the total matched.[19] Formally,

$$\text{Preadjudication IAA} = \text{NMA/NAM}$$

$$\text{Postadjudication IAA} = (\text{NMA} + \text{NEA})\,/\text{NAM}$$

Where NMA is the number of matched mentions, NEA is the number of synonymous mentions, and NAM is the number of annotated mentions.

Preadjudication IAA for all mentions was 67.69%. Postadjudication IAA was 74.20%. Compositional mentions, which account for 19.75% of the total mentions, were more challenging for annotators than mentions that were not compositional. IAA values were 52.21% and 79.61%, respectively.

Evaluation and significance testing

We used accuracy as an evaluation metric and computed statistical significance using approximate randomization.[52,53] As with previous studies,[54–57] we ran approximate randomization using 50 000 shuffles and the significance level set to 0.1.

Systems

In total, 33 teams participated in the 2019 n2c2/UMass Lowell shared task track 3, resulting in 108 total submissions. In our analysis, we focus on the best-performing run of the top 10 performing teams. The methods used by these teams can be broadly divided into 4 groups:

1. Cascading dictionary matching—6 of the top 10 teams applied a series of matching steps of decreasing exactness, starting with exact dictionary matching, followed by hand-crafted rules, edit distance matching, term overlap, or ML-based matching.
2. Cosine distance—one team used the cosine distance between the mention and CUI vectors.
3. Deep learning—one of the participating teams relied on deep learning alone.
4. Retrieve and rank—2 teams performed a retrieval step using IR methods to return a set of candidate CUIs. Then they ranked the candidate CUIs using ML.

Table 2 provides brief descriptions of the top-performing submissions.

Regardless of how CUIs were selected, disambiguation was required by most systems. Ambiguity was often caused by parallel hierarchies of different semantic types in the UMLS. Determining the correct CUI therefore requires inferring the semantic type of mentions. For semantic type prediction, Alibaba (Ali) and ezDI trained deep learning based classifiers. University of Wisconsin–Milwaukee (UWM) trained a traditional ML classifier. Kaiser Permanente (KP) assigned a semantic type based on the observed frequency in the training corpus and the greatest number of supporting vocabularies. Massachusetts Institute of Technology (MIT) used edit distance and hand-crafted rules. University of Arizona (UAZ) used hand-crafted rules. Rather than predict semantic types, Med Data Quest (MDQ)

Table 2. Top performing teams and a brief description of their system

System Type	Team Name	Brief Description
Cascading dictionary matching	ezDI, Inc (ezDI)	Start with exact matches, then hand-crafted text cleaning rules and Levenshtein distance for inexact matches
	Kaiser Permanente (KP)	Start with exact matches, then similarity score based on 3-gram matches
	Massachusetts Institute of Technology (MIT)	Start with exact matches, then a modified edit distance for inexact matches
	National Centre for Text Mining (NaCT)	Start with exact matches, then word overlap, lastly the shortest edit distance
	Toyota Technological Institute of Advanced Industrial Science and Technology	
	University of Arizona (UAZ)	Rule-based ranker for exact matches then character overlap and BioBERT ranker for inexact matches
	University of Wisconsin–Milwaukee (UWM)	Start with exact matches, then learned edit distance rules for inexact matches, and last, submention matching.
Cosine distance	University of Aveiro (UAv)	Cosine similarity between mention and CUI
Deep learning	Toyota Technological Institute (TTI)	Neural network based on cosine similarity between SciBERT and learnable CUI vectors in a dictionary
Retrieve and rank	Alibaba (Ali)	Retrieval using standard IR methods, then ranking using ML
	Med Data Quest, Inc (MDQ)	Retrieval using generated queries, then ranking using ML

The table is sorted alphabetically first by system type, then by team name within each type.
CUI: concept unique identifier; IR: information retrieval; ML: machine learning

disambiguated using the majority type in the training data and the similarity scores between CUI descriptions.

System descriptions

Six teams (ezDI, KP, MIT, National Centre for Text Mining [NaCT], UAZ, UWM) normalized mentions using cascading dictionary matching. Each of these systems started by attempting to find an exact match between the mention and an example in the training data or a CUI in the UMLS. They differed in how inexact matches were handled. For inexact matches KP used a similarity score based on 3-gram matches between the mention and the CUI. UAZ used a BioBERT-based ranker.[17] ezDI applied text cleaning heuristics, used Levenshtein distance, and other manual rules. MIT used a modified edit distance. UWM used a set of automatically learned character-level edit distance rules[58] followed by submention matching. NaCT used Hyphen[59] to clean the data prior to exact matching. For inexact matches, they used the number of shared words and shortest edit distance between the mention and CUI. Notably, UAZ used Lucene (https://lucene.apache.org/) for efficient exact matching and allacronyms.com for acronym expansion.

Toyota Technological Institute (TTI) performed MCN using a deep learning architecture. Interestingly, their architecture was based primarily on the cosine distance between a feature vector of the mention and learnable CUI vectors in the UMLS. The mention vector was constructed by breaking the mention into its subwords (eg, "heart" and "attack" for "heart attack"). The subwords were input into a pretrained SciBERT[60] layer, which was average pooled and input into a fully connected layer. The resulting mention vector was connected to the next layer which computed its cosine distance from learnable CUI vectors for all 434 056 CUIs in SNOMED CT and RxNorm databases (as well as CUI-less). Last, these cosine distances were input into a softmax layer for predicting the CUI for each mention. The system is trained on both the training corpus and the UMLS itself. They used the ArcFace optimization function[61] during training, and tuned hyperparameters using Optuna.[62]

University of Aveiro (UAv) created a cosine distance-based system. They first cleaned the text using hand-crafted rules. Next, they calculated the cosine distance between the embedding of the mention and precalculated CUI embeddings in the UMLS. They used pregenerated biomedical word embeddings.[63] For multiword terms, they used the average of constituent word vectors.

Two teams (Ali, MDQ) created retrieve-and-rank systems. Ali retrieved a set of candidate CUIs using IR methods (eg, frequency, tf-idf, edit distance). Next, they ranked the candidates using ML. The features they used for ML included the output of the IR step, edit distance, dictionary matching, matching atomic unique identifier counts, and semantic types. MDQ built their system using the UIMA[64] framework with Lucene indexing. Candidate CUIs were retrieved via dictionary lookup. Candidates were ranked using ML. The features they used in ML included word embeddings, tf-idf vectors, WordNet, and a BioBERT[17] encoder.

RESULTS

Table 3 presents system performance in ranked order. The first column shows the team name and their rank (based on statistically significant differences in performance), the next column shows the accuracy of the best run of that team, and the next columns show statistical significances. An "X" indicates that there is a statistically significant difference between the 2 teams indicated by the row and column: (1) TTI performs the best and significantly better than all other systems; (2) KP performs significantly better than all but TTI and UAZ; (3) UAZ performs significantly better than all but TTI, KP, and Ali; and (4) Ali, MDQ, and UWM perform better than (5) UAv, ezDI, MIT, and NaCT.

Table 4 presents the aggregate statistics for the best runs among all 33 teams and among the top 10 teams. The best-performing team (TTI) achieved an accuracy 0.8526. The worst-performing system achieved an accuracy of 0.5184. The median was 0.7733, and the mean was 0.7426 with an SD of 0.0858. The differences in performance among the top 10 teams were less. The median, mean, and SD were 0.8090, 0.8111, and 0.0167, respectively. Overall performance was higher than the IAA (74.20%).

Table 3. Accuracy and statistically significant differences between top performing systems.

Team Name (Abbreviation)	Accuracy	Statistical Significances									
		TTI	KP	UAZ	Ali	MDQ	UWM	UAv	ezDI	MIT	NaCT
(1) Toyota Technological Institute (TTI)	0.8526		X	X	X	X	X	X	X	X	X
(2) Kaiser Permanente (KP)	0.8194	X			X	X	X	X	X	X	X
(3) University of Arizona (UAZ)	0.8166	X				X	X	X	X	X	X
(4) Alibaba (Ali)	0.8105	X	X					X	X	X	X
(4) Med Data Quest, Inc (MDQ)	0.8101	X	X	X				X	X	X	X
(4) University of Wisconsin–Milwaukee (UWM)	0.8079	X	X	X				X	X	X	X
(5) University of Aveiro (UAv)	0.8013	X	X	X	X	X	X				
(5) ezDI, Inc (ezDI)	0.8006	X	X	X	X	X	X				
(5) Massachusetts Institute of Technology (MIT)	0.7961	X	X	X	X	X	X				
(5) National Centre for Text Mining (NaCT)	0.7957	X	X	X	X	X	X				

X indicates significant difference between the systems in the row and column.

Table 4. Aggregate statistics over the best runs of the top 10 performing teams

	Among 33 teams	Among top 10 teams
Maximum	0.8526	0.8526
Minimum	0.5184	0.7957
Median	0.7733	0.8090
Mean	0.7426	0.8111
SD	0.0858	0.0167

These results are similar to the results of previous MCN shared tasks. The most directly comparable results are those of ShARe/CLEF 2013 1b and SemEval 2015 Task 2a. These tasks focused on normalization of gold standard *disorder* mentions, and the best-performing teams achieved accuracies of 0.895 and 0.854, respectively. Our best-performing team (TTI) achieved an accuracy of 0.8526 over all mentions. When considering only disorder mentions indicated by the UMLS semantic group "DISO," they remained the best-performing team, and achieved an accuracy of 0.8694 on the 2329 *disorder* mentions in the test set.

Figure 1 shows the average accuracy of the top 10 teams and the percent of mentions in the test set as a function of their length in terms of number of words. A total of 55% of the mentions consisted of single words, and they are correctly normalized 86% of the time by the top systems. Figure 1 indicates that as mention length increases, the number of mentions in the test set decreases, and the average accuracy generally decreases until the mention length reaches 10. This trend of decreasing accuracy as mention length increases reflects the difficulty normalizing complex mentions. Normalizing long mentions (10+ words) is likely easier because there are fewer confounding candidate CUIs.

Because single-word mentions comprise 55% of the test set corpus, they deserve further investigation. Figure 2 shows both the average accuracy and percent of samples in the test corpus vs the character length of the mention. It shows that shorter mentions tend to be more difficult to normalize, and that single character mentions are the most difficult.

To discover why certain systems performed better than others, we also performed a statistical and manual analysis comparing the top systems. Despite differences in system architectures and implementations, most (8 of 10) used very similar normalization techniques based on lexical rules and word and character overlap matching. Notably, TTI and UAv differ from the other systems in that they model linguistic variability with semantic representations.

Figure 3 shows the effect of mention length on accuracy of the top systems. It shows that for all systems but TTI accuracy decreases as mention length increases. TTI's accuracy is unaffected by mention length. Mentions containing more than 11 words are omitted because there are few of them. The distinguishing factor between TTI and other systems is its use of contextual semantic representations (SciBERT). Although UAv used word embeddings, its performance is still affected by mention length. This indicates that contextual semantic representations can overcome the effects of mention length on normalization performance.

Figure 4 shows the accuracy of the top systems on single character mentions. For these, all systems perform similarly poorly, but Ali and ezDI perform noticeably better than others. Ambiguity is particularly problematic for single character mentions, and the distinguishing factor between these 2 systems is that Ali and ezDI performed disambiguation using deep learning with contextual embeddings. These results indicate the importance of incorporating context for normalizing single character mentions. No noticeable differences in performance were found beyond character lengths of 1.

DISCUSSION

Error analysis

A total of 505 mentions were incorrectly predicted by all top 10 systems. We manually analyzed a subset of these to broadly characterize the reasons that they were missed. The 2 main reasons were (1) ambiguous mentions and (2) complex mentions. We further expand on these reasons subsequently and provide characteristic examples of the problems observed. Solving these types of errors will improve performance in future systems.

Ambiguous mentions

Most ambiguities were caused by acronyms, abbreviations, misspellings, and unknown semantic types. Abbreviations and acronyms are much more common in clinical text than in the general domain.[65] Mentions missed by all top systems include examples such as "smear there consistent with ALL," where ALL refers to [*C0023449, Leukemia, Acute Lymphocytic*]. Some of the missed mentions are in lists

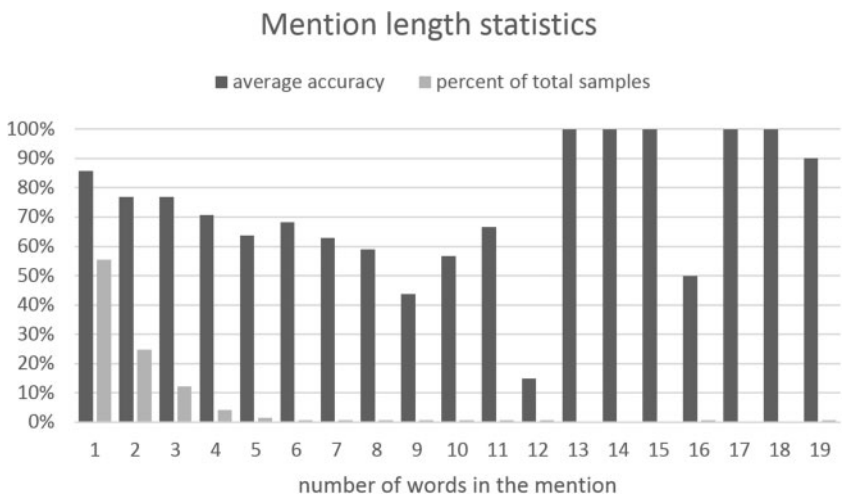

Figure 1. Mention length, in terms of number of words, and average accuracy.

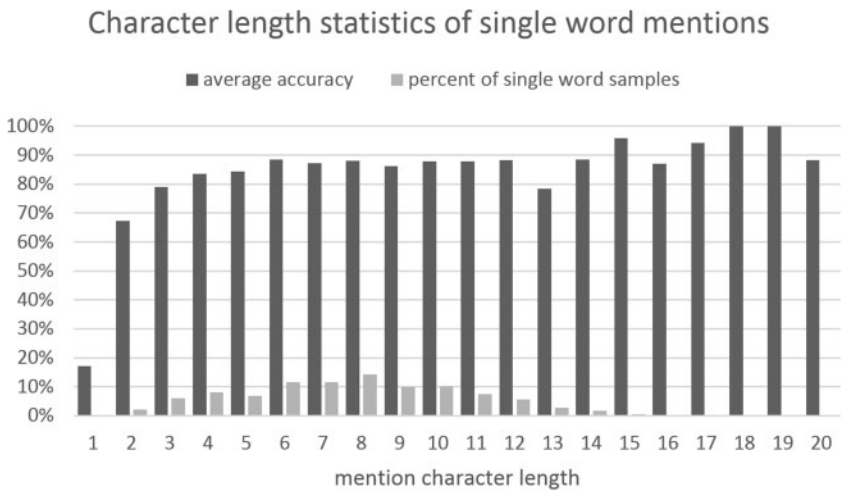

Figure 2. Character length and average accuracy.

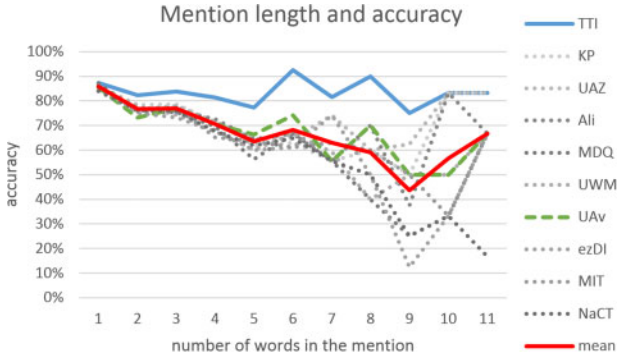

Figure 3. The effects of mention length, in terms of number of words, on the accuracy of semantic and lexical-based systems. Ali: Alibaba; KP: Kaiser Permanente; MDQ: Med Data Quest; MIT: Massachusetts Institute of Technology; NaCT: National Centre for Text Mining; TTI: Toyota Technological Institute; UAZ: University of Arizona; UAv: University of Aveiro; UMW: University of Wisconsin–Milwaukee.

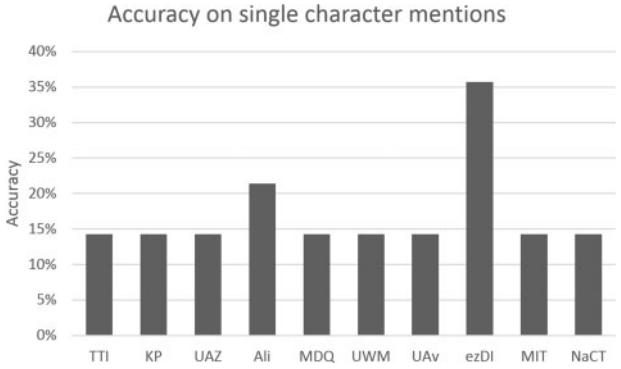

Figure 4. Accuracy of each system on single character mentions. Ali: Alibaba; KP: Kaiser Permanente; MDQ: Med Data Quest; MIT: Massachusetts Institute of Technology; NaCT: National Centre for Text Mining; TTI: Toyota Technological Institute; UAZ: University of Arizona; UAv: University of Aveiro; UMW: University of Wisconsin–Milwaukee.

with many acronyms and abbreviations, such as "no m/r/g PULM—CTAB no w/r/r ABD -nt/nd; incision c/d/I," where nd maps to [*C0577599, swelling is absent*]. Lists like this are common for clinical records, and future systems may require acronym dictionaries, knowledge of the context, and knowledge of clinical note structure to normalize them. This difficulty with abbreviations echoes findings from ShARe/CLEF 2013. There, task 2 focused exclusively on normalizing abbreviations whereas task 1b focused on normalizing disorders. The best results for task 2 were 0.719 vs 0.895 for task 1b. These qualitative results are supported by the results in Figure 2, as it shows that terms with few characters such as acronyms and abbreviations, and particularly single character terms are exceptionally difficult to normalize.

Misspelled mentions were also problematic, particularly misspelled brand names: examples are "acuchecks" for accuchecks, a brand name for [*C0430032, Glucometer blood glucose*], and "Duracef" for Duricef, a brand name for [*C0007538, Cephadroxyl*].

Additionally, because mentions can map to different CUIs in parallel SNOMED CT hierarchies, there were many errors where knowing the semantic type was necessary. Although most teams incorporated semantic type prediction into their algorithms, it was not perfect. An example missed by all top teams was "positive pressure at the bedside," in which bedside should map to [*C0558274, Bed area*] but was most often incorrectly mapped to [*C0282662, bedside testing*].

Complex mentions

Complex mentions typically contained many words, were very specific, or required domain knowledge to normalize. Complex mentions should not to be confused with compositional annotations. Compositional annotations require more than 1 CUI to describe the annotation, whereas complex mentions map to a single CUI. Although ambiguity played a role in errors related to complex mentions, the root problem was that systems were unable to fully represent and understand them. Nevertheless, system's incorrect normalizations were often closely related to the correct ones. Incorrect normalization as CUI-less was also common.

As an example, consider "there was a mild diminution of light touch, pinprick, position, and vibration sense the left side," for [*C0020580, Hypesthesias (decreased sense of touch/sensation)*]. In this case, the mention is a list of symptoms. Incorrect normalizations for this mention included [*C1285608, observable sense of touch*] and [*C1295585, decreased vibratory sense (finding)*]. Another characteristic mention is "to the left anterior descending artery and diag-

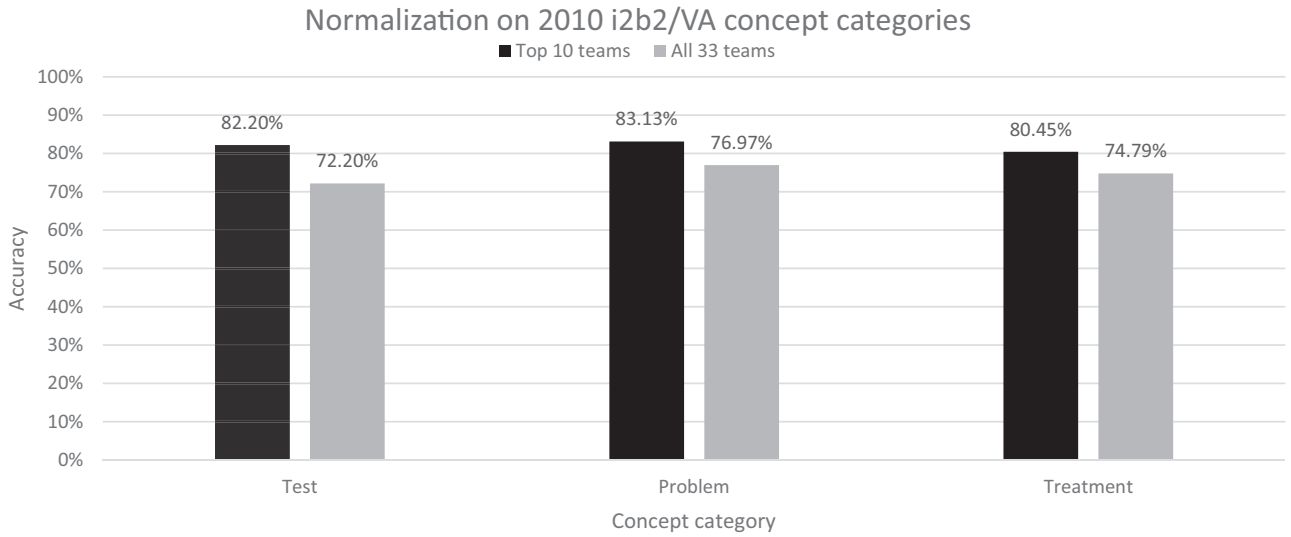

Figure 5. Normalization on 2010 i2b2/VA concept categories.

onal" for [*C0226034, structure of distal portion of anterior descending branch of left coronary artery*]. This mention includes a list of modifiers specifying the type and location of "artery." The most common incorrect prediction for this mention was [*C0226032, structure of anterior descending branch of left coronary artery*]. For these types of mentions, heuristics such as term overlap and vector-based representations built from the sum or average of constituent words may be inappropriate. Future systems may require more complex mention representations and mapping techniques. For examples such as "diminution of light touch, pinprick, position, and vibration sense the left side," a system may first need to recognize that the mention is describing symptoms. It may then utilize domain knowledge such as UMLS glosses for normalization. For the "left anterior descending artery and diagonal" example, a parse tree could help select the most appropriate CUI or help navigate the UMLS hierarchy based on how words in the mention relate to "artery." Alternatively, a mapping could potentially be improved if UMLS is better represented by learned embeddings that capture the local node structure around each CUI.

Other cases are even more complex, and likely require handcrafted rules. For example: "a Gleason 's IV, plus V tumor," should be mapped to [*C0332334, Gleason grade score 9 out of 10*], but was incorrectly mapped to [*C0332329, gleason grade 4*]. A correct mapping of this mention requires both domain knowledge that a Gleason grade of IV plus Gleason grade of V equals a Gleason grade of 9, and understanding that the mention implies that the 2 scores should be summed.

Concept category

In addition to the overall performance evaluation, we calculated the normalization accuracy for each of the three 2010 i2b2/VA concept categories included in the MCN corpus: problems, treatments, and tests. For compositional annotations in the 2010 i2b2/VA corpus, we measured the weighted accuracy over the component mentions. For example, "bowel wall thickening," in 2010 i2b2/VA annotation was split into 2 subsumed concept mentions, "bowel wall" and "thickening," in the MCN annotation. In this case, each correct normalization contributes 0.5 score to the accuracy. We report the average normalization scores for each category in Figure 5. The results showed similar normalization accuracy among 3 categories. This indicates, importantly, that widening the scope for normalized clinical concepts in fact does not make the task more challenging for current state-of-the-art concept normalization methods. At the same time, the MCN strategy of splitting compositional mentions into subspans[19] may have simplified the normalization task for difficult cases, making the overall task more accessible for automated systems.

CONCLUSIONS

MCN acounts for variation in term usage by mapping mentions to a controlled vocabulary. For this shared task, 33 teams submitted system runs. Among the top 10 performing teams, we observed 4 primary architectures: (1) cascading dictionary matching, (2) cosine distance, (3) deep learning, and (4) retrieve and rank. Overall, systems performed well and exceeded IAA; the median accuracy of all participating teams was 0.7733, and the IAA was 0.7420. The best-performing system was a deep learning system that used the cosine distance between SciBERT embeddings and learnable CUI vectors. It achieved an accuracy of 0.8526.

MCN is a challenging task, even for human annotators due to ambiguities and linguistic complexities of mentions. Ambiguous mentions include acronyms, abbreviations, misspellings, and unknown semantic types. More accurate spelling correction, better acronym and abbreviation resolution, and better semantic type prediction can help MCN of these mentions. Complex mentions include long multiword expressions, which are often very specific. Normalizing these mentions likely requires the development of more complex representations. Furthermore, such representations may help systems to recognize and automatically break up compositional mentions (eg, "left breast biopsy," or "breast or ovarian cancer"), which required manual standardization during annotation. Future work should focus on these ambiguous and complex cases, in which exact dictionary matching, word overlaps, edit distance, and simple vector representations fail. As a starting point, the analysis indicates that contextual semantic representations help overcome challenges associated with long mentions and single character mentions.

FUNDING

Y-FL and AR were supported in part by a research grant from Philips HealthCare. SH and OU were supported by the National Library of Medicine of the National Institutes of Health grant numbers R13LM013127 (to OU) and R13LM011411 (to OU).

AUTHOR CONTRIBUTIONS

Y-FL and AR organized and ran the shared task. Y-FL and SH performed data analysis and error analysis. SH is the primary author of this manuscript, with contributions from OU, Y-FL, and AR. SH, OU, YW, and FS organized the n2c2 workshop at which the shared task results were presented. Y-FL and SH contributed equally to this work.

ACKNOWLEDGEMENTS

We thank Weiyi Su for her contribution to the development of the MCN corpus.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. MIT Critical Data. *Secondary Analysis of Electronic Health Records*. New York, NY: Springer Nature; 2016.
2. Dalianis H. *Clinical Text Mining: Secondary Use of Electronic Patient Records*. New York, NY: Springer Nature; 2018.
3. Shickel B, Tighe PJ, Bihorac A, *et al.* Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J Biomed Health Inform* 2017; 22 (5): 1589–604.
4. Singh Gangwar P, Hasija Y. Deep learning for analysis of Electronic Health Records (EHR). *Deep Learning Techniques for Biomedical and Health Informatics*. Cham, Switzerland: Springer; 2020: 149–66.
5. Miotto R, Li L, Kidd BA, *et al.* Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep* 2016; 6 (1): 26094–10.
6. Barroilhet SA, Pellegrini AM, McCoy TH, *et al.* Characterizing DSM-5 and ICD-11 personality disorder features in psychiatric inpatients at scale using electronic health records. *Psychol Med* 2020; 50 (13): 2221–9.
7. Zhou S-M, Fernandez-Gutierrez F, Kennedy J, *et al.*; UK Biobank Follow-up and Outcomes Group. Defining disease phenotypes in primary care electronic health records by a machine learning approach: a case study in identifying rheumatoid arthritis. *PLoS One* 2016; 11 (5): e0154515.
8. Nguyen BP, Pham HN, Tran H, *et al.* Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. *Comput Methods Programs Biomed* 2019; 182: 105055.
9. Ye C, Fu T, Hao S, *et al.* Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning. *J Med Internet Res* 2018; 20 (1): e22.
10. Zheng T, Xie W, Xu L, *et al.* A machine learning-based framework to identify type 2 diabetes through electronic health records. *Int J Med Inform* 2017; 97: 120–7.
11. Rumshisky A, Ghassemi M, Naumann T, *et al.* Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Transl Psychiatry* 2016; 6 (10): e921.
12. Sabra S, Mahmood Malik K, Alobaidi M. Prediction of venous thromboembolism using semantic and sentiment analyses of clinical narratives. *Comput Biol Med* 2018; 94: 1–10.
13. Liu R, Greenstein JL, Sarma SV, *et al.* Natural language processing of clinical notes for improved early prediction of septic shock in the ICU. *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*; 2019: 6103–8.
14. Buchan K, Filannino M, Uzuner Ö. Automatic prediction of coronary artery disease from clinical narratives. *J Biomed Inform* 2017; 72: 23–32.
15. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004; 32(suppl_1): D267–70.
16. Devlin J, Chang M-W, Lee K, *et al.* BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*; 2019: 1.
17. Lee J, Yoon W, Kim S, *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020; 36 (4): 1234–40.
18. Alsentzer E, Murphy J, Boag, W, *et al.* Publicly available clinical BERT embeddings. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*; 2019: 72–8.
19. Luo Y-F, Sun W, Rumshisky A. MCN: a comprehensive corpus for medical concept normalization. *J Biomed Inform* 2019; 92: 103132.
20. Uzuner Ö, South BR, Shen S, *et al.* 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011; 18 (5): 552–6.
21. Spackman KA, Campbell KE, Côté RA. SNOMED RT: a reference terminology for health care. *Proc AMIA Annu Fall Symp* 1997; 1997: 640–4.
22. Liu S, Ma W, Moore R, Ganesan V, Nelson S. Rxnorm: prescription for electronic drug information exchange. *IT Professional* 2005; 7 (5): 17–23.
23. Luo Y-F, Sun W, Rumshisky A. A hybrid normalization method for medical concepts in clinical narrative using semantic matching. *AMIA Jt Summits Transl Sci Proc* 2019; 2019: 732–40.
24. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001; 2001: 17.
25. 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.
26. 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.
27. Leaman R, Khare R, Lu Z. Challenges in clinical natural language processing for automated disorder normalization. *J Biomed Inform* 2015; 57: 28–37.
28. Suominen H, Salanterä S, Velupillai S, *et al.* Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. *International Conference of the Cross-Language Evaluation Forum for European Languages*. New York, NY: Springer; 2013: 212–31.
29. Pradhan S, Elhadad N, Chapman W, Manandhar S, Savova G. Semeval-2014 task 7: analysis of clinical text. In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*; 2014: 54–62.
30. Elhadad N, Pradhan S, Gorman S, Manandhar S, Chapman W, Savova G. Semeval-2015 task 14: Analysis of clinical text. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*; 2015: 303–10.
31. Hirschman L, Colosimo M, Morgan A, Yeh A. Overview of BioCreative task 1b: normalized gene lists. *BMC Bioinformatics* 2005; 6 (Suppl 1): S11.
32. Morgan AA, Lu Z, Wang X, *et al.* Overview of BioCreative II gene normalization. *Genome Biol* 2008; 9 (Suppl 2): S3.
33. Lu Z, Kao H-Y, Wei C-H, *et al.* The gene normalization task in BioCreative III. *BMC Bioinformatics* 2011; 12 (S8): S2.
34. Li J, Sun Y, Johnson RJ, *et al.* BioCreative v CDR task corpus: a resource for chemical disease relation extraction. *Database (Oxford)* 2016; 2016: baw068.
35. Roberts K, Demner-Fushman D, Tonning JM. Overview of the TAC 2017 adverse reaction extraction from drug labels track. *Proceedings of Text Analysis Conference (TAC) 2017*; 2017.
36. Leaman R, Miller C, Gonzalez G. Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark. *Proc 2009 Symp Lang Biol Med* 2009; 82 (9).

37. 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.
38. Sarker A, Gonzalez-Hernandez G. Overview of the second social media mining for health (SMM4H) shared tasks at AMIA 2017. *Training* 2017; 822 (10): 1239.
39. Limsopatham N, Collier N. Adapting phrase-based machine translation to normalise medical terms in social media messages. arXiv, doi: https://arxiv.org/abs/1508.02285, 10 Aug 2015, preprint: not peer reviewed.
40. Limsopatham N, Collier N. Normalising medical concepts in social media texts by learning semantic representation. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*; 2016: 1014–23.
41. Karimi S, Metke-Jimenez A, Kemp M, Wang C. CADEC: a corpus of adverse drug event annotations. *J Biomed Inform* 2015; 55: 73–81.
42. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res* 2010; 39(suppl_1): D52–57.
43. Lipscomb CE. Medical subject headings (MeSH). *Bull Med Libr Assoc* 2000; 88 (3): 265–6.
44. Brown EG, Wood L, Wood S. The medical dictionary for regulatory activities (MEDDRA). *Drug Saf* 1999; 20 (2): 109–17.
45. Fescharek R, Kübler J, Elsasser U, Frank M, Güthlein P. Medical dictionary for regulatory activities (MEDDRA). *Int J Pharm Med* 2004; 18 (5): 259–69.
46. Davis AP, Wiegers TC, Rosenstein MC, Mattingly CJ. Medic: a practical disease vocabulary used at the comparative toxicogenomics database. *Database (Oxford)* 2012; 2012: bar065.
47. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005; 33 (Database issue): D514–17.
48. Kuhn M, Letunic I, Jensen LJ, Bork P. The sider database of drugs and side effects. *Nucleic Acids Res* 2016; 44 (D1): D1075–79.
49. NEHTA. Australian Medicines Terminology v3 Model – Common – Release Note v1.4. 2014. https://developer.digitalhealth.gov.au/specifications/ehealth-foundations/ep-1825-2014/nehta-1827-2014. Accessed December 1, 2020.
50. Saeed M, Lieu C, Raber G, Mark RG. MIMIC II: a massive temporal icu patient database to support research in intelligent patient monitoring. *Compututers in Cardiology* 2002; 29: 641–4.
51. Stubbs A. MAE and MAI: lightweight annotation and adjudication tools. *Proceedings of the 5th Linguistic Annotation Workshop*; 2011: 129–33.
52. Noreen EW. *Computer-Intensive Methods for Testing Hypotheses*. New York, NY: Wiley; 1989.
53. Yeh A. More accurate tests for the statistical significance of result differences. *Proceedings of the 18th Conference on Computational linguistics-Volume 2*; 2000: 947–53.
54. 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.
55. Chinchor N. The statistical significance of the MUC-4 results. *Proceedings of the 4th Conference on Message Understanding*; 1992: 30–50.
56. Stubbs A, Filannino M, Soysal E, Henry S, Uzuner Ö. Cohort selection for clinical trials: n2c2 2018 shared task track 1. *J Am Med Inform Assoc* 2019; 26 (11): 1163–71.
57. Stubbs A, Kotfila C, Uzuner Ö. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task track 1. *J Biomed Informatics* 2015; 58: S11–9.
58. Kate RJ. Normalizing clinical terms using learned edit distance patterns. *J Am Med Inform Assoc* 2016; 23 (2): 380–6.
59. Thompson P, Ananiadou S. HYPHEN: a flexible, hybrid method to map phenotype concept mentions to terminological resources. *Terminology* 2018; 24 (1): 91–121.
60. Beltagy I, Cohan A, Lo K. SciBERT: pretrained contextualized embeddings for scientific text. arXiv, doi: https://arxiv.org/abs/1903.10676, 10 Sep 2019.
61. Deng J, Guo J, Xue N, Zafeiriou S. Arcface: Additive angular margin loss for deep face recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2019: 4690–9.
62. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; 2019: 2623–31.
63. Chen Q, Peng Y, Lu Z. BioSentVec: creating sentence embeddings for biomedical texts. arXiv, doi: https://arxiv.org/abs/1810.09302, 24 Jan 2020
64. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Nat Lang Eng* 2004; 10 (3–4): 327–48.
65. Moon S, Pakhomov S, Melton GB. Automated disambiguation of acronyms and abbreviations in clinical texts: window and training size considerations. *AMIA Annu Symp Proc* 2012; 2012: 1310–9.