Advancing the State of the Art in Clinical Natural Language Processing through Shared Tasks

Michele Filannino[1,2], Özlem Uzuner[1,2]
[1] George Mason University, Fairfax, VA, USA
[2] Massachusetts Institute of Technology, Cambridge, MA, USA

Summary

Objectives: To review the latest scientific challenges organized in clinical Natural Language Processing (NLP) by highlighting the tasks, the most effective methodologies used, the data, and the sharing strategies.

Methods: We harvested the literature by using Google Scholar and PubMed Central to retrieve all shared tasks organized since 2015 on clinical NLP problems on English data.

Results: We surveyed 17 shared tasks. We grouped the data into four types (synthetic, drug labels, social data, and clinical data) which are correlated with size and sensitivity. We found named entity recognition and classification to be the most common tasks. Most of the methods used to tackle the shared tasks have been data-driven. There is homogeneity in the methods used to tackle the named entity recognition tasks, while more diverse solutions are investigated for relation extraction, multi-class classification, and information retrieval problems.

Conclusions: There is a clear trend in using data-driven methods to tackle problems in clinical NLP. The availability of more and varied data from different institutions will undoubtedly lead to bigger advances in the field, for the benefit of healthcare as a whole.

Keywords

Clinical natural language processing; shared tasks; scientific challenges; survey

Yearb Med Inform 2018:184-92
http://dx.doi.org/10.1055/s-0038-1667079

1 Introduction

Recent years have seen an increase in the number of scientific challenges, also called shared tasks, organized for the advancement of Natural Language Processing (NLP) in clinical data [1]. Shared tasks promote work specific to a "challenge question" posed to the research community and aim to evaluate the state of the art. Without the unifying framework of a shared task, even though the NLP community might work on the same general problem, the nuances of problems will vary, to the point where the approaches would not be comparable. For this reason, shared task organizers often provide a data set, annotated with gold standard annotations, for system development and tuning. The evaluation of the systems on the challenge question takes place on a held-out data set. This setup provides a way of comparing systems head-to-head on the same data and task, and helps identify the state of the art. The shared task data may remain available for research beyond the challenge time frame, providing a common benchmark for assessing the quality of future attempts [2]. They also provide a great resource for training future generations: they are a great instrument to advance the research and engage students. The ready availability of datasets, evaluation scripts, and commentary provides an ideal environment that serves as a catalyst and motivator.

In addition to the above benefits, in the clinical domain, because of the scarcity of data and their poor availability, shared tasks make it possible for the global research community to tackle problems that would otherwise be inaccessible to them [3]. However, attaining these benefits requires overcoming some obstacles [4]:

- Availability of data: Clinical data, i.e., data that contains clinical information, can take many forms. Most often used synonymously with electronic health records (EHRs), clinical data can contain social media data as well as information from resources such as drug labels. Each of these forms of data comes with its own challenges for access and use
- De-identification: Health Insurance Portability and Accountability Act (HIPAA) [5] defines requirements for safeguarding of patient health data, indicating elements of private health information (PHI) that need to be protected. De-identification, i.e., removal of PHI from records, provides one way of addressing this concern. However, there is a downside to de-identification: this process alters the contents of the original records and as a result some useful information may be lost. On the other hand, HIPAA-compliant de-identification may not be adequate in some cases, e.g., professions, which are not covered by HIPAA, are allowed to remain in the records even though, when rare enough, they could uniquely identify patients. This makes de-identification a challenging process that needs to strike a delicate balance between de-identifying the data, so that it can be shared, and preserving the medical content of the data, so that it can be useful for downstream medical research. As a result, de-identification often requires a manual review of the data – an expensive and time-consuming process that ultimately limits the size of the shared data
- Annotation: Often the bigger cost in shared task organization comes from gold standard generation for the clinically-relevant task that is posed to the community

[6]. Gold standard generation requires input from experts who are well-versed in the tasks studied. Experts tend to be medical professionals with high hourly rates – another parameter that needs to be balanced against the volume of desired data.

In this paper, we review the latest scientific challenges organized to tackle NLP problems on clinical data. We highlight the tasks and the most effective methodologies used to tackle these tasks, along with the data used.

2 Methods

This review focuses on shared tasks using clinical data to tackle NLP problems. The relevant studies have been identified by querying Google Scholar and PubMed with "((shared task) OR challenge) AND (clinical OR health OR EHR) AND (NLP)". The returned articles were limited to those describing clinical NLP shared tasks which were published since 2015. This resulted in a total of 17 shared tasks. Four challenges took place in 2015, six in 2016 and seven in 2017. Sixteen shared tasks are complete and published. One is completed but still in the process of publishing the outcomes. For a survey of shared tasks held before 2015, one can refer to Velupillai et al. [7]. For a survey in the broader field of biomedical text mining, one can refer to Huang et al. [8].

3 Shared Tasks

Recent clinical NLP shared tasks have utilized social media data (e.g., Twitter, forum posts), journal articles (e.g., MEDLINE/PubMed), as well as electronic health records (e.g., pathology reports, nursing admission notes, psychiatric evaluation records, etc.) and other health-related documents such as drug labels. Collectively, these shared tasks posed questions on a variety of data, including both de-identified real data and synthetic records. Table 1 summarizes the key characteristics of each of the shared tasks. We present the shared tasks according to the type of data they use.

3.1 Synthetic Data

Synthetic data can serve as a placeholder for real data and allows to side-step the privacy issues related to real data. The downside of synthetic data is that its generation comes with a cost and must make sure that the synthetic data captures the characteristics of real data so that the solutions developed can be valid on real data.

The 2015 Text REtrieval Conference (TREC) Clinical Decision Support (CDS) shared task aimed at evaluating biomedical retrieval systems [9]. The organizers provided a set of 30 synthetic case narratives (called topics), consisting of a short textual description, a summary, and a diagnosis. They asked the participants to develop systems for retrieving the most relevant scientific articles within a collection of 733,138 articles[1] from PubMed Central (PMC)[2]. Thirty-six teams participated in this task, 33 from academia, three from industry. The top performing system achieved an inferred normalized discounted cumulative gain (infNDCG) of 38.21% [10] by combining several Information Retrieval (IR) models (BM25, PL2, BB2).

The 2017 TREC Precision Medicine (PM) shared task [11] utilized 30 semi-structured synthetic topics (e.g., disease, genetic variants, demographic information, and other factors) and evaluated IR systems for their ability to match topics with: 1) 26,759,399 abstracts from MEDLINE; and 2) 241,006 clinical trial descriptions from ClinicalTrials.gov[3]. Thirty-two teams participated in this task, 27 from academia, five from industry. The top performing system achieved a precision at 10 (P@10) of 63.10% and 44.29% for track 1 and 2, respectively. This system combined a query expansion module with a heuristic scoring method for abstracts and trials.

The Conference and Labs of the Evaluation Forum (CLEF) eHealth 2016[4] shared task [12] used the National Information and Communications Technology Australia (NICTA) Synthetic Nursing Handover Data [15]. This data set consisted of 300 notes that were authored by a registered nurse[5]. Each note consisted of a patient profile and a free-form text paragraph. One of the proposed tasks asked to the participants on this data was to automatically pre-populate handover forms with relevant text-snippets (slot filling) [16]. Three teams participated in this task, all of them from academia. The top performing system scored 38.2% (F1-score) and relied on a Conditional Random Field (CRF) model that used a set of features extracted from Stanford CoreNLP, Unified Medical Language System (UMLS) [17], WordNet, regular expression patterns, and Latent Dirichlet Allocation (LDA) clusters [18]. A wrapper algorithm evaluated several different subsets of these features and ultimately selected the best one.

3.2 Real Data

Prescription Drug Labels

Prescription drug labels published by the Food and Drug Administration (FDA) contain information about uses of medications, indications, and side effects. They are meant for public use and are free of any privacy concern[6]. This makes them a good target for studying medication-related problems, such as identifying adverse drug reactions (ADRs), comparing ADRs presented in labels from different manufacturers for the same drug, and performing pharmacovigilance by identifying new ADRs not currently included in labels.

The 2017 Text Analysis Conference (TAC) ADR Extraction from Drug Labels [19] studied FDA drug labels. The organizers shared a dataset of 2,309 unannotated drug

[1] Available for download at http://trec-cds.appspot.com/2015.html

[2] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[3] Available for download at http://trec-cds.appspot.com/2017.html

[4] Although CLEF eHealth is organized every year [13, 14], its main focus is multi-linguality and information retrieval rather than clinical NLP

[5] Available for download at https://sites.google.com/site/clefehealth2016/

[6] Accessible at http://www.drugs-library.com

Category	Year	Challenge name	Task description	Data type	Data source	teams type			
						Academia	Industry	Joint	Total
Synthetic	2015	TREC Clinical Decision Support (CDS) [9]	Patient-centered information retrieval	Medical case narratives	Synthetic, PubMed	33	3	0	36
	2017	TREC Precision Medicine [11] > Track 1 > Track 2	Patient-centered literature article retrieval Patient-centered clinical trials retrieval	Semi-structured cases	Synthetic, PubMed, ClinicalTrials.gov	27	5	0	32
	2016	CLEF eHealth [12]	Information extraction	Nursing handover notes	NICTA synthetic nursing handover notes	4	0	0	4
Prescription drug labels	2017	Text Analysis Conference (TAC) Adverse Drug Reaction Extraction from Drug Labels (ADR) [18] > Track 1 > Track 2 > Track 3 > Track 4	ADR mentions and modifiers extraction Relation extraction Positive ADR filtering Positive ADR normalization	Drug labels	Drugs-Library.com	6	3	1	10
Online social data	2015	CLPsych: Depression and PTSD on Twitter [22]	Binary classification of depression and PTSD users	Social media	Twitter	3	0	0	3
	2016	Social Media Mining (SMM) [24] > Track 1 > Track 2 > Track 3	ADR classification Information extraction Concept normalization	Social media	Twitter	9	2	0	11
	2017	Social Media Mining for Health Applications (SMM4HA) [29] > Track 1 > Track 2 > Track 3	ADR classification Classification of medication intake Concept normalization	Social media	Twitter	12	1	0	13
	2016	CLPsych: Triaging content in online peer-support forums [33]	Classification of mental health severity in 4 levels	Forum	ReachOut	13	1	1	15
	2017	CLPsych: Triaging content in online peer-support forums [35]	Classification of mental health severity in 4 levels	Forum	ReachOut	12	2	1	15
	2017	NTCIR-13 MedWeb [36]	8-class classification of diseases and symptoms	Multilingual Social media	Twitter	7	1	1	9
Clinical data	2015	Analysis of Clinical Text (ACT) [39] > Track 1 > Track 2a > Track 2b	Disorder NER and normalization Template slot filling (given gold spans) Disorder recognition and template slot filling (end-to-end)	Clinical notes	ShARe corpus (MIMIC)	18	3	0	21
	2016	TREC Clinical Decision Support (CDS) [43]	Patient-centered IR	Nursing admission notes	MIMIC, PubMed	21	5	0	26
	2017	Medication and Adverse Drug Events (MADE1.0) > Track 1 > Track 2	Medication, ADE, sign and symptom identification Relation extraction	Clinical notes	UMass Memorial Medical Center	-	-	-	-
	2015	Clinical TempEval [45] > Track 1 > Track 2 > Track 3	Time expression extraction Event extraction Relation extraction (wrt DCT) Relation extraction (wrt narrative containers)	Pathology reports	Mayo Clinic	3	0	0	3
	2016	Clinical TempEval [46] > Track 1 > Track 2 > Track 3	Time expression extraction Event extraction Relation extraction (wrt DCT) Relation extraction (wrt narrative containers)	Pathology reports	Mayo Clinic	11	3	0	14
	2017	Clinical TempEval [48] > Track 1 > Track 2 > Track 3	Time expression extraction (cross-domain) Event extraction (cross-domain) Relation extraction (wrt DCT) Relation extraction (wrt narrative containers)	Pathology reports, Clinical notes	Mayo Clinic	9	2	0	11
	2016	Centers for Excellence in Genomics N-GRID (CEGS-NGRID) [51] > Track 1a > Track 1b > Track 2	De-identification (cross-domain) De-identification Psychiatric Symptom Severity Prediction	Psychiatric evaluation records	Partners Healthcare and Harvard Medical School	23	5	3	31

Table 1 Clinical NLP Challenges, the tasks they posed, and the number of participating teams, since 2015, ordered by data sensitivity.

labels, 200 of which manually annotated with ADR spans, relations, and concept identifiers (IDs)[7]. TAC proposed four tasks: 1) ADR mentions and modifiers span extraction; 2) extraction of relations between ADRs and their corollaries; 3) filtering of positive ADRs; and 4) positive ADR normalization [20]. Ten teams took part in this task, six from academia, three from industry, and one joint team. The same system ranked first on all tasks, where it achieved an F1-score of 82.48%, 49.00%, 82.19% (macro F1), and 85.33% (macro F1), respectively. This system used two distinct bi-directional Long Short Term Memory (LSTM) -CRF models with some post-processing rules to tackle the first two tasks. A learning-to-rank approach using RankSVM (support vector machine) on the top 10 normalization candidates tackled Tasks 3 and 4.

Online Social Data

Among the information shared in social media are personal views, experiences, and even health information [21]. However, social media data are not free of privacy and ethics concerns [22]. Access to most social media data requires a registration and consent to the governing rules, which can prevent secondary uses and limit the maximum amount of data to be collected. If social media data are not de-identified, then they cannot be shared among institutions and must be (re-)obtained directly from their source, e.g., Twitter data are often "distributed" in the form of tweet IDs, user IDs, and download scripts. Since 2015, there have been six clinical NLP shared tasks that used social media data. Four of them have been manually de-identified (or anonymized) and require a data use agreement (DUA) to be signed. Some are available for download beyond the challenges' timeframes.

The 2015 Computational Linguistics and Clinical Psychology Workshop (CLPsych) used Twitter data for classifying users based on depression and post-traumatic stress disorder (PTSD) [23]. The organizers collected, anonymized, and annotated tweets of the form "I have just been diagnosed with X", with "X" being depression or PTSD. The resulting dataset included 7,857 million tweets from 477 depression patients, 396 PTSD patients, and 1,746 control users. The data were distributed according to Twitter terms of service, along with a privacy agreement that required protective measures for downloaded copies. The data are available for download[8] and require Institutional Review Board (IRB) approval and signing of the privacy policy. Four teams participated in this task, three from academia, one from industry. The best performing system achieved an average precision above 80% [24] and was based on a Support Vector Machine (SVM) with linear kernel and baseline lexical features with term-frequency-inverse document frequency (TF-IDF) weighting.

The 2016 Social Media Mining shared task (SMM) [25] studied tweets for identifying ADRs. A data set of 10,822 anonymized tweets [26] was annotated by two pharmacology experts and was made available to the participants[9]. The shared task consisted of three tracks: 1) classification of tweets as ADR- and non-ADR-related; 2) ADR span extraction from tweets; and 3) linking ADRs to their UMLS [17] concepts. Eleven teams took part in this task but only six are reported in the overview: four from academia, two from industry. In the first track, the best performing system achieved an F1-score of 41.95% [27] by using an ensemble of Random Forest models with unigram, bigram, and trigram features. Track 2 was tackled as a Named Entity Recognition (NER) task by all the participants with the most effective machine learning (ML) model being CRFs and achieving 61.10% F1-score [28] on a subset of the entire corpus (2,131 annotated tweets). The organizers did not receive submissions for Track 3. Track 1 was re-proposed at the 2017 workshop [29] along with two new tasks: classification of medication intake types, and normalization of clinical concepts to the Medical Dictionary for Regulatory Activities (MedDRA) [20]. The 2017 workshop also extended the 2016 dataset to 15,717 tweets for training and 9,961 for testing. For classification of ADR-related tweets, the top performing system achieved an F1-score of 43.5% with an SVM model trained on textual features and domain-specific word embeddings [30]. For classification of medication intake, the top performing system scored F1 at 69.3% and used convolutional neural networks (CNNs) on word embeddings [31]. Finally, the top performing system for concept normalization scored an F1-score of 88.5% and used an ensemble of linear and deep learning models [32].

The 2016 CLPsych shared task [33] used 65,024 posts from the online forum of ReachOut, an Australian non-profit that supports young people. A total of 1,227 posts were manually prioritized by three independent judges by how urgently they need a response from a moderator (i.e., paraprofessional support) in a 4-point scale. The remaining posts were left un-annotated to experiment with semi-supervised and unsupervised techniques. Fifteen teams took part in this task: 13 from academia, one from industry, and one joint team. The top performing system achieved a macro-averaged F1-score of 42% by using an ensemble of classifiers working on different granularity of text [34]. The task was repeated in 2017 with an expanded dataset (157,963 posts, of which 1,588 were annotated[10]) and attracted a similar number of teams [35]. The best performing team obtained a macro-averaged F1-score of 46.7%. The data from 2016 and for 2017 are available for download on request.

Finally, the 2017 NII Testbeds and Community for Information access and Research'13 (NTCIR-13) MedWeb shared task[11] used a dataset of 2,560 tweets in Japanese, English, and Chinese [36]. The organizers manually de-identified the data and shared them with the participants under a DUA. Participants were asked to label the data with eight diseases/symptoms: influenza, diarrhea, hay fever, cough/sore throat, headache, fever, runny nose, and cold. Four academic teams took part in the English subtask by submitting 12 systems. The best system [37] achieved an exact match accuracy of 88% by using an ensemble of hierarchical attention

[7] Available for download at https://bionlp.nlm.nih.gov/tac2017adversereactions/

[8] http://www.cs.jhu.edu/~mdredze/datasets/clpsych_shared_task_2015/

[9] Can be downloaded using a script at http://diego.asu.edu/downloads

[10] http://clpsych.org/shared-task-2017/

[11] http://mednlp.jp/medweb/NTCIR-13/

networks (HAN) and deep character-level convolutional neural networks (CNNs). At the time of writing, only the training data was available for download[12].

Clinical Notes

Clinical notes constitute the most sensitive set of data for shared tasks. They are governed by HIPAA and access to these data can require human subjects training, as well as DUAs even when they are de-identified. Medical Information Mart for Intensive Care (MIMIC) is the most frequently used source of de-identified clinical notes. It contains health data of over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012 [38]. Since 2015, two shared tasks have utilized MIMIC as their data set. Other shared tasks have used de-identified and annotated data from their own home institutions. Unless noted otherwise, these data are distributed with DUAs.

The 2015 Analysis of Clinical Text (ACT) shared task [39] utilized the ShaRe dataset [40] consisting of 531 manually annotated discharge summaries, electrocardiograms, echo, and radiology reports from MIMIC-II. ACT focused on two tasks: 1) detection and normalization of disorder mentions; and 2) template slot filling. Twenty-one teams took part in this task, 18 from academia, three from industry. These teams tackled the first task as a sequence labelling problem, using CRFs in combination with word embeddings and ad-hoc sentence clustering. The second task was proposed in two settings according to whether the participants used the gold or the predicted spans for the disorder mentions (Track 2.a and 2.b respectively). The best performing system for the first task scored 75.7% (strict F1-score) [41]. On the second task, the same system scored first in both settings: 88.6% (weighted accuracy score) on Track 2.a and 80.8% (F1 * weighted accuracy score) on Track 2.b [42]. This system tackled the tasks by using a combination of CRFs and a binary SVM, both based on part-of-speech tags and syntactic features.

The 2016 TREC Clinical Decision Support shared tasks [43] studied patient-centered IR. The organizers provided a set of 1.25 million scientific articles from PubMed Central (PMC), and 30 nursing admission notes from MIMIC-III (called topics). With permission from the MIMIC team, the notes were made publicly available without the need for a DUA. Even though the notes were already de-identified, the de-identification process was manually carried out a second time for maximum privacy protection. For consistency with the previous challenges in the series [9, 44],, only the notes' history of present illness sections were provided to the participants[13]. Participants were asked to retrieve articles relevant for answering questions on diagnoses, tests, and treatments. Twenty-six teams took part in the challenge: 21 from academia, and five from industry. The top performing system achieved a precision at 10 of 40.33%, which is higher than the best score achieved in 2015 (see above). Despite this, the average results were lower than in 2015. The organizers ascribed the result to the difference of real Intensive Care Unit (ICU) notes from synthetic general practice notes.

NLP Challenges for Detecting Medication and Adverse Drug Events from Electronic Health Records (MADE1.0)[14] utilized 1,092 medical records from 21 cancer patients in the UMass Memorial Medical Center to propose three tasks: 1) clinical named entity recognition; 2) relation identification; and 3) end-to-end systems to conduct the first two tasks together. This shared task is currently completed but the overview paper is not published yet.

The Clinical TempEval challenges [45, 46], hosted at the Semantic Evaluation series (SemEval), used 600 de-identified clinical notes and pathology reports from cancer patients at the Mayo Clinic that are manually annotated with temporal expressions, medical events, and temporal relations[15]. In 2015, three academic teams participated in this shared task. In 2016, 14 teams participated, three of which were from industry. The 2016 results were better than those of the previous year, with the top performing system in time expression extraction achieving an F1-score of 79.5% [47] using linear and structural SVMs on morphological, syntactic, discourse, and word representation features. The same system ranked first also in the event (F1-score 90.3%) and temporal relations tasks (F1 75.6% for the relations with respect to document creation times (DCTs), and F1 47.9% for the ones among narrative containers).

Clinical TempEval 2017 [48] studied a domain adaptation problem, from colon cancer to brain cancer pathology reports and clinical notes. The corpus for this task contained 1,216 notes from each of the two types of cancer patients at the Mayo Clinic[16]. The notes were manually de-identified and annotated by experts [6][17]. Eleven teams took part in this shared task: nine from academia, and two from industry. The best performing system achieved F1-scores of 57% for time expression spans (using an ensemble of CRFs, rules and decision trees) [49], 72% for event spans, 59% for temporal relations with respect to the DCT, and 32% for those among narrative containers [50]. The system used neural networks with character and word embeddings combined with SVMs. Those results were approximately 20% lower than the ones registered by systems trained and tested on the same domain [45, 46].

Finally, the CEGS-NGRID Shared Tasks and Workshop on Challenges in NLP for Clinical Data made available a corpus of 1,000 manually de-identified psychiatric evaluation records from Partners Healthcare [51]. The organizers extended the HIPAA definition of PHI for better privacy protection. They proposed two tasks: 1) de-identification [52], and 2) symptom severity prediction [53]. De-identification was studied in two subtasks: a) benchmarking pre-existing de-identification systems [54, 55] on psychiatric records[18] (called "sight unseen"); and b) regular de-identification. Overall, 31 teams took part, 23 from academia, five from

[12] http://mednlp.jp/medweb/NTCIR-13/#dataset

[13] Available for download at http://trec-cds.appspot.com/2016.html

[14] https://bio-nlp.org/index.php/projects/39-nlp-challenges

[15] The data are available under DUA at http://alt.qcri.org/semeval2015/task6/ (2015) and http://alt.qcri.org/semeval2016/task12/ (2016).

[16] The data are available under DUA at http://alt.qcri.org/semeval2017/task12/

[17] The annotations are available at https://github.com/stylerw/thymedata

[18] Only unannotated test data were released to prevent participants from adapting systems to the new data.

industry, and three jointly from industry/academia. The same system scored the highest in both subtasks of Task 1: F1-score of 79.85% [52], and F1-score 91.43% [55] respectively. The system used a combination of CRFs, BI-LSTMs, and rules. The result suggests that "out-of-the-box solutions provide a good start at building models that can be tuned to the new data". In the second task, symptom severity prediction, the systems were scored using the Inverse Normalized Mean Absolute Error Macro-averaged ($INMAE^M$), which weights a prediction's error according to its ordinal distance from the correct class. The top performing system used an ensemble of machine learning classifiers based on morphological, syntactic, and structural features and achieved an $INMAE^M$ score of 86.3% which is close to the level of accuracy recorded by the least experienced of the annotators.

The information presented in this section highlights how the data varies in its sensitivity to privacy, which inversely correlates with the available data size. Tasks range from NER to relation extraction, multi-class classification problems and information retrieval, with these last ones being the most successful in terms of both attracting participation and system performance. The use of CRF and BI-LSTM models is common to almost all the top performing NER systems. More diverse methods are used for the relation extraction and multi-class classification problems.

4 Discussion

The discussed shared tasks offer interesting insights related to the availability of data, the advances in the state-of-the-art techniques, the role of privacy, and the importance of data size in supporting the methodological advances.

4.1 Data Availability

The concerns of availability of data, privacy, and cost of annotation ultimately shape the landscape of the field and give direction to the state of the art. Attempts to bypass concerns of availability of data and privacy with synthetic data results in displacing the cost of de-identification to the cost of generating synthetic records and come at the risk of generating a synthetic set that may not represent real data perfectly. Efforts to use social media data to understand the user perspective on her/his health problems face the same kind of privacy concerns as the notoriously sensitive EHR data. They additionally run into constraints related to long term access to data: either they do not remain available after the challenge or they need to be re-obtained from the social media site itself. When the data are to be re-obtained, this leaves the fate of the data set in the hands of the users of social media and could be lost if the users delete the messages or their accounts.

4.2 Observing Advances in the State of the Art

Shared tasks continue to grow both in their numbers and in the participation they attract. Especially for the tasks that are organized regularly, the consistency in the tasks and growing datasets continue to attract growing numbers of participants. Some tasks such as de-identification and NER tend to recur because of their high practical value. Table 2 shows the performances of the systems participating in the most recent shared tasks. It shows that tasks such as clinical named entity recognition (medications, times, events, PHI) are well understood with system performances above 70% (see TAC ADR 2017, CLPsych 2015, ACT 2015, and CEGS-NGRID 2016), while tasks such as relation extraction with performances below 50% need more attention (see the Clinical TempEval series). Clinical information retrieval tasks, with a performance around 50% (see the TREC series), show the need for further research. Finally, multi-class classification tasks (see the CLPsych series) show a performance below 50%, which can be partly justified by the lack of annotated data.

4.3 Balancing Access, Privacy, and Corporate Confidentiality

Interestingly, until now, academic institutions have dominated shared task participation. Few of the shared tasks reviewed in this paper had a significant participation from industry (e.g., the TREC series and CEGS N-GRID). Industry bridges the gap between pure research and technology [56]. However, the stringent rules governing the use of data and the hesitation to openly share the methods for fear of losing intellectual property result in decreased participation. Attracting more companies to shared tasks would help in diversifying the methods and contributions, reduce the gap between academia and industry, and shorten the time it takes for methods to be adopted by industry.

DUAs required from participants before access to data vary in complexity. Some DUAs pose really strict requirements, e.g., storing the data on machines that are not connected to the Internet for the entire duration of the challenge. Keeping the terms of DUAs to those requirements that match the sensitivity level of data could open up more data sets to more parties for research and encourage participation of more parties.

4.4 Larger Datasets Support Methodological Advances

The approaches used to tackle problems in clinical NLP are almost entirely in the realm of data-driven methods. Named entity recognition tasks, such as medication or ADR extraction, are commonly solved using CRFs or deep learning approaches (BI-LSTMs), often with word embeddings although n-gram features are still used. Classification and relation extraction tasks are tackled using ensembles, often as a way of coping with the imbalance nature of classes. This makes a compelling argument for advocating the adoption of bigger datasets. Despite increasing the cost of design and annotation, richer data sets have the benefit of increasing the external validity of the developed solutions.

5 Conclusions

In this paper we reviewed the latest scientific challenges organized in clinical NLP, by highlighting the tasks, the most effective methodologies used, the data, and the sharing strategies. We surveyed 17 shared tasks, grouped by the type of data used (synthetic,

Category	Year	Challenge name	Task description	Data type	Data source	Data size	De-identification / anonymization	DUA	Currently Available?	Best Performance	Measure
Synthetic	2015	TREC Clinical Decision Support (CDS) [9]	Patient-centered information retrieval	Medical case narratives	Synthetic, PubMed	30 topics, 730K articles	no	no	yes	38.21%	infNDCG
	2017	TREC Precision Medicine [11]		Semi-structured cases	Synthetic, PubMed, ClinicalTrials.gov	30 topics, 27M abstracts, 241K trials	no	no	yes		
		> Track 1	Patient-centered literature article retrieval							63.10%	P@10
		> Track 2	Patient-centered clinical trials retrieval							44.29%	P@10
	2016	CLEF eHealth [12]	Information extraction	Nursing handover notes	NICTA synthetic nursing handover notes	300 notes	no	no	yes	38.20%	F1 (macro avg.)
Prescription drug labels		Text Analysis Conference (TAC) Adverse Drug Reaction Extraction from Drug Labels (ADR) [18]									
		> Track 1	ADR mentions and modifiers extraction							82.48%	F1
	2017	> Track 2	Relation extraction	Drug labels	Drugs-Library.com	2309 labels	no	no	yes	49.00%	F1
		> Track 3	Positive ADR filtering							82.19%	F1 (macro avg.)
		> Track 4	Positive ADR normalization							85.33%	F1 (macro avg.)
Online social data	2015	CLPsych: Depression and PTSD on Twitter [22]	Binary classification of depression and PTSD users	Social media	Twitter	7.8M tweets	yes	yes	yes	80.00%	Avg. Precision
	2016	Social Media Mining (SMM) [24]		Social media	Twitter	10,882 tweets	no	no	yes		
		> Track 1	ADR classification							41.95%	F1
		> Track 2	Information extraction							61,10%	F1
		> Track 3	Concept normalization							-	
	2017	Social Media Mining for Health Applications (SMM4HA) [29]		Social media	Twitter	15,777 tweets	no	no	yes		
		> Track 1	ADR classification							43.50%	F1
		> Track 2	Classification of medication intake							69.30%	F1 (micro avg.)
		> Track 3	Concept normalization							88.50%	Accuracy
	2016	CLPsych: Triaging content in online peer-support forums [33]	Classification of mental health severity in 4 levels	Forum	ReachOut	65,024 (1,227 annotated)	yes	yes	yes, on request	42.00%	F1 (macro avg.)
	2017	CLPsych: Triaging content in online peer-support forums [35]	Classification of mental health severity in 4 levels	Forum	ReachOut	157,963 posts (1,588 annotated)	yes	yes	yes, on request	46.70%	F1 (macro avg.)
	2017	NTCIR-13 MedWeb [36]	8-class classification of diseases and symptoms	Multilingual Social media	Twitter	2560 tweets	yes	yes	yes, on request	-	
Clinical data	2015	Analysis of Clinical Text (ACT) [39]		Clinical notes	ShARe corpus (MIMIC)	531 summaries	yes	yes	yes		
		> Track 1	Disorder NER and normalization							75.70%	F1 (strict)
		> Track 2a	Template slot filling (given gold spans)							88.60%	F1 * weighted acc.
		> Track 2b	Disorder recognition and template slot filling (end-to-end)							80.80%	F1 * weighted acc.
	2016	TREC Clinical Decision Support (CDS) [43]	Patient-centered IR	Nursing admission notes	MIMIC, PubMed	30 notes, 1.25M abstracts	yes	no	yes	40.33%	P@10
	2017	Medication and Adverse Drug Events (MADE1.0)		Clinical notes	UMass Memorial Medical Center	1092 records	yes	yes	no		
		> Track 1	Medication, ADE, sign and symptom identification							-	
		> Track 2	Relation extraction								
	2015	Clinical TempEval [45]		Pathology reports	Mayo Clinic	600 notes	yes	yes	yes, on request		
		> Track 1	Time expression extraction							72.50%	F1
		> Track 2	Event extraction							87.50%	F1
		> Track 3	Relation extraction (wrt DCT)							70.20%	F1
			Relation extraction (wrt narrative containers)							12.30%	F1
	2016	Clinical TempEval [46]		Pathology reports	Mayo Clinic	600 notes	yes	yes	yes, on request		
		> Track 1	Time expression extraction							79.50%	F1
		> Track 2	Event extraction							90.30%	F1
		> Track 3	Relation extraction (wrt DCT)							75.60%	F1
			Relation extraction (wrt narrative containers)							47.90%	F1
	2017	Clinical TempEval [48]		Pathology reports, Clinical notes	Mayo Clinic	1216 notes	yes	yes	yes, on request		
		> Track 1	Time expression extraction (cross-domain)							57.00%	F1
		> Track 2	Event extraction (cross-domain)							72.00%	F1
		> Track 3	Relation extraction (wrt DCT)							59.00%	F1
			Relation extraction (wrt narrative containers)							32.50%	F1
	2016	Centers for Excellence in Genomics N-GRID (CEGS-NGRID) [51]		Psychiatric evaluation records	Partners Healthcare and Harvard Medical School	1000 records	yes	yes	yes, on request		
		> Track 1a	De-identification (cross-domain)							79.85%	F1
		> Track 1b	De-identification							91.43%	F1
		> Track 2	Psychiatric Symptom Severity Prediction							86.30%	INMAE^M

Table 2 List of shared tasks with data source, data size, sub-tasks descriptions, and best-performance score (metrics differ per challenge). The table also contains information about data availability after the challenge, whether the data have been de-identified, and whether they require a DUA to be signed.

drug labels, social data, and clinical data). We found that the type of data is correlated with its size and sensitivity. Recognition and classification of named entities are the most common tasks, usually tackled by data-driven approaches.

We hope that the growing number of success stories in shared task organization will encourage more institutions to share data. More and varied data from different institutions will undoubtedly lead to bigger advances in the field, for the benefit of healthcare as a whole.

Acknowledgments
We wish to thank Pierre Zweigenbaum, Aurélie Névéol, and the anonymous reviewers for their valuable comments.

References

1. Ohno-Machado L. Realizing the full potential of electronic health records: the role of natural language processing. J Am Med Inform Assoc 2011 Sep 1;18(5):539.
2. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner Ö. 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 Sep;18(5):540-3.
3. Nissim M, Abzianidze L, Evang K, van der Goot R, Haagsma H, Plank B, et al. Last Words: Sharing is Caring: The Future of Shared Tasks. Computational Linguistics 2017;43(4):897-904.
4. Lluch M. Healthcare professionals' organizational barriers to health information technologies — A literature review. Int J Med Inform 2011 Dec 31;80(12):849-62.
5. Dwyer SJ 3rd, Weaver AC, Hughes KK. Health insurance portability and accountability act. Security Issues in the Digital Medical Enterprise 2004 Apr;72(2):9-18.
6. Styler WF 4th, Bethard S, Finan S, Palmer M, Pradhan S, de Groen PC, et al. Temporal annotation in the clinical domain. Trans Assoc Comput Linguist 2014 Apr 30;2:143-54. (http://aclweb.org/anthology/Q/Q14/Q14-1012.pdf)
7. Velupillai S, Mowery D, South BR, Kvist M, Dalianis H. Recent advances in clinical natural language processing in support of semantic analysis. Yearb Med Inform 2015;10(1):183-93.
8. Huang CC, Lu Z. Community challenges in Biomedical Text Mining over 10 years: success, failure and the future. Brief Bioinform 2015 May 1;17(1):132-44.
9. Roberts K, Simpson MS, Voorhees EM, Hersh WR. Overview of the TREC 2015 Clinical Decision Support Track. In: Proceedings of the 2015 Text Retrieval Conference.
10. Song Y, He Y, Hu Q, He L. Ecnu at 2015 CDS track: Two re-ranking methods in medical information retrieval. In: Proceedings of the 2015 Text Retrieval Conference 2015.
11. Roberts K, Demner-Fushman D, Voorhees EM, Hersh WR, Bedrick S, Lazar AJ. Overview of the TREC 2017 precision medicine track. TREC, Gaithersburg, MD; 2017.
12. Kelly L, Goeuriot L, Suominen H, Névéol A, Palotti J, Zuccon G. Overview of the CLEF eHealth evaluation lab 2016. In International Conference of the Cross-Language Evaluation Forum for European Languages 2016 Sep 5. Springer International Publishing; 2016. P. 255-66.
13. Goeuriot L, Kelly L, Suominen H, Hanlen L, Névéol A, Grouin C, et al. Overview of the CLEF eHealth evaluation lab 2015. In: International Conference of the Cross-Language Evaluation Forum for European Languages 2015 Sep 8. Springer, Cham, 2015. p. 429-43.
14. Goeuriot L, Kelly L, Suominen H, Névéol A, Robert A, Kanoulas E, et al. CLEF 2017 eHealth evaluation lab overview. In: International Conference of the Cross-Language Evaluation Forum for European Languages 2017 Sep 11. Springer, Cham; 2017. p. 291-303.
15. Suominen H, Zhou L, Hanlen L, Ferraro G. Benchmarking clinical speech recognition and information extraction: new data, methods, and evaluations. JMIR Med Inform 2015 Apr;3(2):e19.
16. Suominen H, Zhou L, Goeuriot L, Kelly L. Task 1 of the CLEF eHealth Evaluation Lab 2016: Handover Information Extraction. In CLEF (Working Notes) 2016 Sep. p. 1-14.
17. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 1;32(Database issue):D267-70.
18. Ebersbach M, Herms R, Lohr C, Eibl M. Wrappers for Feature Subset Selection in CRF-based Clinical Information Extraction. In CLEF (Working Notes) 2016; p. 69-80.
19. Roberts K, Demner-Fushman D, Tonning JM. Overview of the TAC 2017 Adverse Reaction Extraction from Drug Labels Track. Proceedings of the Text Analysis Conference; 2017.
20. Brown EG, Wood L, Wood S. The medical dictionary for regulatory activities (MedDRA). Drug Saf 1999 Feb 1;20(2):109-17.
21. Gross R, Acquisti A. Information revelation and privacy in online social networks. In: Proceedings of the 2005 ACM workshop on Privacy in the electronic society 2005 Nov 7. ACM; 2005. p 71-80).
22. Zimmer M. "But the data is already public": on the ethics of research in Facebook. Ethics Inf Technol 2010 Dec 1;12(4):313-25.
23. Coppersmith G, Dredze M, Harman C, Hollingshead K, Mitchell M. CLPsych 2015 Shared Task: Depression and PTSD on Twitter. In: CLPsych@ HLT-NAACL 2015 Jun 5. p. 31-9. (http://www.aclweb.org/anthology/W15-1204)
24. Resnik P, Armstrong W, Claudino L, Nguyen T. The University of Maryland CLPsych 2015 shared task system. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality 2015. p. 54-60. (http://www.aclweb.org/anthology/W15-1207)
25. Sarker A, Nikfarjam A, Gonzalez G. Social Media Mining shared task workshop. In: Biocomputing 2016: Proceedings of the Pacific Symposium 2016. p. 581-92.
26. Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, et al. Utilizing social media data for pharmacovigilance: A review. J Biomed Inform 2015 Apr 30;54:202-12.
27. Rastegar-Mojarad MA, Elayavilli RK, Yu Y, Liu H. Detecting signals in noisy data-can ensemble classifiers help identify adverse drug reaction in tweets. In: Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing 2016.
28. Wang CK, Singh ON, Dai HJ, Jonnagaddala JI, Jue TR, Iqbal US, et al. NTTMUNSW system for adverse drug reactions extraction in Twitter data. In Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing, Big Island, HI, USA 2016 Jan. p. 4-8.
29. Sarker A, Gonzalez-Hernandez G. Overview of the Second Social Media Mining for Health (SMM4H) Shared Tasks at AMIA 2017. In: Proceedings of the 2nd Social Media Mining for Health Research and Applications Workshop;1(10,822):1239.
30. Kiritchenko S, Mohammad SM, Jason Morin JC, de Bruijn B. NRC-Canada at SMM4H Shared Task: Classifying Tweets Mentioning Adverse Drug Reactions and Medication Intake. In: Proceedings of the Second Workshop on Social Media Mining for Health Applications (SMM4H). Health Language Processing Laboratory; 2017.
31. Friedrichs J, Mahata D, Gupta S. InfyNLP at SMM4H Task 2: Stacked Ensemble of Shallow Convolutional Neural Networks for Identifying Personal Medication Intake from Twitter. In: Proceedings of the Second Workshop on Social Media Mining for Health Applications (SMM4H). Health Language Processing Laboratory; 2017.
32. Belousov M, Dixon W, Nenadic G. Using an Ensemble of Generalised Linear and Deep Learning Models in the SMM4H 2017 Medical Concept Normalisation Task. In: Proceedings of the Second Workshop on Social Media Mining for Health Applications (SMM4H). Health Language Processing Laboratory; 2017.
33. Milne DN, Pink G, Hachey B, Calvo RA. CLPsych 2016 Shared Task: Triaging content in online peer-support forums. In CLPsych@ HLT-NAACL 2016. p. 118-27. (http://www.aclweb.org/anthology/W16-0312)
34. Mac Kim S, Wang Y, Wan S, Paris C. Data61-CSIRO systems at the CLPsych 2016 Shared Task. In CLPsych@ HLT-NAACL 2016. p. 128-32. (http://www.aclweb.org/anthology/W16-0313)
35. Hollingshead K, Ireland ME, Loveys K. Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology–From Linguistic Signal to Clinical Reality; 2017.
36. Wakamiya S, Morita M, Kano Y, Ohkuma T, Aramaki E. Overview of the NTCIR-13: Medweb task. In Proceedings of the NTCIR-13 Conference; 2017.
37. Iso H, Ruiz C, Murayama T, Taguchi K, Takeuchi R, Yamamoto H,et al. NTCIR-13 MedWeb Task: Multi-label Classification of Tweets using an Ensemble of Neural Networks. In Proceedings of the NTCIR-13 Conference 2017.
38. Saeed M, Villarroel M, Reisner AT, Clifford G.

Lehman LW, Moody G, et al. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database. Crit Care Med 2011 May;39(5):952.
39. Elhadad N, Pradhan S, Gorman SL, Manandhar S, Chapman WW, Savova GK. SemEval-2015 Task 14: Analysis of Clinical Text. In SemEval@ NAACL-HLT 2015 Jun 4 (pp. 303-310). (http://aclweb.org/anthology/S/S15/S15-2051.pdf)
40. Mowery DL, Velupillai S, South BR, Christensen L, Martinez D, Kelly L, et al. Task 2: ShARe/CLEF eHealth evaluation lab 2014. In: Proceedings of CLEF: 2014.
41. Pathak P, Patel P, Panchal V, Soni S, Dani K, Patel A, Choudhary N. ezDI: A Supervised NLP System for Clinical Narrative Analysis. In: SemEval@ NAACL-HLT 2015 Jun 4. p. 412-6. (http://aclweb.org/anthology/S/S15/S15-2071.pdf)
42. Xu J, Zhang Y, Wang J, Wu Y, Jiang M, Soysal E, et al. UTH-CCB: The Participation of the SemEval 2015 Challenge-Task 14. In: SemEval@NAACL-HLT 2015 Jun 4. p. 311-4. (http://aclweb.org/anthology/S/S15/S15-2052.pdf)
43. Roberts K, Demner-Fushman D, Voorhees E, Hersh W. Overview of the TREC 2016 Clinical Decision Support Track. In: Proceedings of the Twenty-Five Text REtrieval Conference (TREC 2016), Nov 2016, Gaithersburg, United States.
44. Simpson MS, Voorhees EM, Hersh W. Overview of the TREC 2014 Clinical Decision Support Track. In: Proceedings of the 2014 Text Retrieval Conference.
45. Bethard S, Derczynski L, Savova G, Pustejovsky J, Verhagen M. SemEval-2015 Task 6: Clinical TempEval. In: SemEval@NAACL-HLT 2015 Jun 4. p. 806-14. (http://aclweb.org/anthology/S/S15/S15-2136.pdf)
46. Bethard S, Savova G, Chen WT, Derczynski L, Pustejovsky J, Verhagen M. Semeval-2016 Task 12: Clinical TempEval. Proceedings of the 10th International Workshop on Semantic Evaluations (SemEval-2016); 2016. p. 1052-62. (http://www.aclweb.org/anthology/S16-1165)
47. Lee HJ, Xu H, Wang J, Zhang Y, Moon S, Xu J, et al. UTHealth at SemEval-2016 Task 12: an End-to-End System for Temporal Information Extraction from Clinical Notes. In: SemEval@ NAACL-HLT 2016. p. 1292-17. (http://www.aclweb.org/anthology/S16-1201)
48. Bethard S, Savova G, Palmer M, Pustejovsky J. SemEval-2017 Task 12: Clinical TempEval. Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017); 2017. p. 565-72. (http://aclweb.org/anthology/S17-2000)
49. MacAvaney S, Cohan A, Goharian N. GUIR at SemEval-2017 Task 12: A Framework for Cross-Domain Clinical Temporal Information Extraction. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) 2017. p. 1024-9. (http://www.aclweb.org/anthology/S17-2180)
50. Tourille J, Ferret O, Névéol A, Tannier X. LIMSI-COT at SemEval-2016 Task 12: Temporal relation identification using a pipeline of classifiers. In: SemEval@NAACL-HLT 2016. p. 1136-42. (http://www.aclweb.org/anthology/S16-1175)
51. Uzuner Ö, Stubbs A, Filannino M. A natural language processing challenge for clinical records: Research Domains Criteria (RDoC) for psychiatry. J Biomed Inform 2017 Oct 16.75:S1-S3. (https://doi.org/10.1016/j.jbi.2017.10.005)
52. Stubbs A, Filannino M, Uzuner,Ö. De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID Shared Tasks Track 1. J Biomed Inform 2017 Nov;75S:S4-S18.
53. Filannino M, Stubbs A, Uzuner Ö. Symptom severity prediction from neuropsychiatric clinical records: Overview of 2016 CEGS N-GRID Shared Tasks Track 2. J Biomed Inform 2017 Nov;75S:S62-S70.
54. Uzuner Ö, Luo Y, Szolovits,P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc 2007 Sep 1;14(5):550-63.
55. Jiang Z, Zhao C, He B, Guan Y, Jiang J. De-identification of medical records using Conditional Random Fields and Long Short-Term Memory networks. J Biomed Inform 2017 Nov;75S:S43-S53.
56. Clements D, Dault M, Priest A. Effective teamwork in healthcare: research and reality. Healthc Pap 2007;7(I):26.

Correspondence to:
Özlem Uzuner
4400 University Drive, MS 1G8
5359 Nguyen Engineering Building
Fairfax, VA 22030, USA
Tel: +1 703 993 5996
E-mail: ouzuner@gmu.edu