Published in final edited form as:

Drug Saf. 2019 January ; 42(1): 99–111. doi:10.1007/s40264-018-0762-z.

Overview of the First Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE 1.0)

Abhyuday Jagannatha¹, Feifan Liu², Weisong Liu^{3,4}, Hong Yu^{1,3,4,5,*}

¹College of Information and Computer Sciences, University of Massachusetts, Amherst

²Department of Quantitative Health Sciences and Radiology, University of Massachusetts Medical School, Worcester

³Department of Computer Science, University of Massachusetts, Lowell

⁴Department of Medicine, University of Massachusetts Medical School, Worcester

⁵Bedford VAMC, Bedford

Abstract

Introduction—This work describes the MADE 1.0 corpus and provides an overview of the MADE 2018 challenge for Extracting Medication, Indication and Adverse Drug Events from Electronic Health Record Notes.

Objective—The goal of MADE is to provide a set of common evaluation tasks to assess the state of the art for NLP systems applied to electronic health records (EHRs) supporting drug safety surveillance and pharmacovigilance. We also provide benchmarks on the MADE dataset using the system submissions received in MADE 2018 challenge.

Methods—The MADE 1.0 challenge has released an expert-annotated cohort of medication and adverse drug event information, comprised of 1,089 fully de-identified longitudinal EHR notes from 21 randomly selected cancer patients at the University of Massachusetts Memorial Hospital. Using this cohort as a benchmark, the MADE 1.0 challenge designed three shared NLP tasks. The named entity recognition (NER) task identifies medications and their attributes (dosage, route, duration, and frequency), indications, adverse drug events (ADEs) and severity. The relation identification (RI) task identifies relations between the named entities: medication-indication, medication-ADE, and attribute relations. The third shared task (NER-RI) evaluates NLP models

* hong.yu@umassmed.edu.

Compliance with Ethical Standards

Declaration

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Dataset

The data used is from the MADE 1.0 corpus available at bio-nlp.org/dataset/made1

Conflicts of Interests

The authors do not have any conflicts of interest to declare.

that perform the NER and RI tasks jointly. Eleven teams from four countries participated in at least one of the three shared tasks and forty-one system submissions were received in total.

Results—The best systems f-scores for NER, RI, and NER-RI are 0.82, 0.86, and 0.61 respectively. Ensemble classifiers using the team submissions improved the performance further, with an f-score of 0.85, 0.87 and 0.66 for the three tasks respectively

Conclusion—MADE results show that recent progress in NLP has led to remarkable improvements in NER and RI tasks for the clinical domain. However, there is still some room for improvement, particularly in the NER-RI task.

1 Introduction

An adverse drug event (ADE) is “an injury resulting from a medical intervention related to a drug” [1]. ADEs are the single largest contributor to hospital-related complications in inpatient settings [2] and comprise approximately one-third of all hospital adverse effects (AE). They affect more than two million hospital stays annually [3] and prolong hospital length of stay by 1.7 to 4.6 days [4, 5]. These events also account for approximately two-thirds of all post-discharge complications, more than a quarter of which are estimated to be preventable [6]. National estimates suggest that ADEs contribute at least an additional 30 billion dollars to U.S. health care costs [7].

ADEs should ideally be detected by the randomized control trial (RCT) studies before the relevant drug ever enters the market. However, due to a limited number of participants and inclusion/exclusion criteria reflecting specific subject characteristics (demographic, medical condition and diagnosis, age) [8] pre-marketing RCT frequently miss ADEs. This assertion is supported by the fact that FDA withdrawal rate of once approved drugs over the first sixteen years ranges from 21% to 27% [9]. Drug safety surveillance and post-marketing pharmacovigilance (PhV), “the science and activities relating to the detection, assessment, understanding, and prevention of adverse effects or any other drug-related problem,”[10] are therefore vitally important tools for monitoring FDA-approved drug safety.

One of the earliest systems for post-marketing PhV aimed at improving the drug safety are spontaneous reporting systems (SRSs) such as the FDA Adverse Event Reporting System (FAERS), a voluntary SRS. Although SRSs have been highly successful for PhV, they have limitations such as under-reporting [11, 12] and missing important patterns of drug exposure [13]. To counter these shortcomings, other resources have been proposed for PhV, including biomedical literature [14] and social media [15]. However, biomedical literature has been shown to identify only a limited set of ADEs, mainly rare ADEs [16]. Social media also has challenges, such as incomplete and erroneous drug exposure patterns and duplication [17].

It is well-known that electronic health records (EHRs) contain rich ADE information and have been widely used for drug safety surveillance and PhV [2, 6]. Unlike other resources which are passive in nature, EHRs can be a rich resource for real-time or active PhV and patient-drug surveillance. In addition, they can lead to better and more cost-effective, patient management [18]. In 2009, the FDA initiated the mini-sentinel program to facilitate the use of routinely collected EHR data for active surveillance of marketed medical product safety

[19]. For example, Yih et al. [20] have shown an increased risk of intussusception after rotavirus vaccination in US infants.

However, most of the EHR-based PhV and patient safety surveillance systems are based on the analyses of the structured data such as ICD codes [21, 22]. It is well-known that ADEs are often buried in the clinical narrative [23–25] and are not separately recorded in diagnosis codes or other structured data fields. Even information that is expected to be reported in the structured fields, such as bodyweight, frequently appears only in EHR text [26]. In addition, necessary information that can be used to assess the causality of a medication and adverse drug event, including temporal and causal relations, only exist in the narratives. However, extraction of ADE information from EHR narratives remains a challenge because manual data extraction is very costly. It is, therefore, a significant impediment to large-scale PhV studies.

Natural Language Processing(NLP) may be a solution to provide fast, accurate and automated ADE detection that can yield significant cost and logistical advantages over the aforementioned practices of manual chart review or voluntary reporting [27]. However, despite the advances in NLP, few methods are specifically developed for detecting ADE information from EHR notes. Several NLP systems use resources like UMLS [28] to extract disease and drug mentions to generate ADE predictions using co-occurrence metrics. Such approaches may miss other important drug information (e.g., indication and dose), and would fail to capture temporal and/or causal associations between a drug and an ADE that may be explicitly expressed in EHR narratives. Moreover, different NLP systems have been evaluated on different gold standards, making it challenging to identify state of the art NLP technologies.

Therefore, we have created the MADE 1.0 corpus, a publicly available, expert-curated benchmark of EHR notes that have been annotated with clinical named entities (i.e., drug name, dosage, route, duration, frequency, indication, ADE, and other signs and symptoms) and relations (ADE-Drugname, Indication-Drugname, Drugname-attributes etc.). The MADE corpus is the first dataset that provides detailed annotations for Medication, Indication, Adverse Drug Events, their attributes and relations relevant to drug safety surveillance and PhV studies.

Using this high-quality corpus as a benchmark, we designed three shared tasks (NER, RI, and NER-RI) to assess the state of the art NLP technologies which have the potential to improve downstream PhV related tasks. These shared tasks were organized in the First Natural Language Processing Challenge for Detecting Medication and Adverse Drug Events from Electronic Health Records (MADE). This challenge was hosted by the University of Massachusetts (Amherst, Lowell, and Worcester) from August 2017 to March 2018. In this paper we first describe the MADE corpus. We then document the shared tasks and provide a comprehensive report of system submissions in the MADE challenge. The main contributions of this paper are as follows.

- To present the first richly annotated and publicly available EHR data for adverse drug events detection and drug surveillance research.

- To describe the carefully designed schema and release the detailed annotation guidelines, which would be a valuable resource to not only drug safety but also any other data-driven clinical informatics research.
- To introduce three shared tasks in the MADE challenge and report the system submissions and results.
- To perform an ensemble-based system aggregation which shows that the top systems are complementary and can be further integrated to push the boundaries of extracting medication, indication, and ADEs from EHRs.

2 Related Work

Natural language processing (NLP) techniques have been widely applied to biomedicine [28–33]. Much of NLP research in the biomedical domain has centered on named entity recognition and normalization tasks. Examples of shared tasks in this domain include BioNLP [34], BioCreAtivE [35], i2b2 shared NLP tasks [36] and ShARe/CLEF evaluation tasks [37].

Existing NLP approaches for EHR ADE detection can be grouped into rule-based, lexicon-based, supervised machine learning, and hybrid approaches. For example, Li et al. [38] built an NLP system with the knowledge of a domain expert. Melton and Hripcsak [39] applied the NLP system MedLEE, a rule-based semantic parser, to detect concepts. Similarly, Humphreys et al. [40] applied MedLEE to map free text to the Unified Medical Language System (UMLS) concepts and semantic types. UMLS [41] is a resource which combines multiple biomedical and clinical resources into a unified ontology. Rochefort et al. [42] developed supervised machine learning classifiers to classify whether a clinical note contains deep venous thromboembolisms (DVT) and pulmonary embolism (PE) using bag-of-words from EHR narratives. Haerian et al. [43] applied distance supervision to identify terms (e.g., including “suicidal”, “self-harm,” and “diphenhydramine overdose”) associated with assigned suicide ICD 9 code, and then used those terms to recover suicide events. Wang et al. [44] used MetaMap to identify drugs mentioned in the text threads of online health forums. Nikfarjam et al. [45] annotated adverse drug event information on user posts from Daily-Strength and Twitter. They then used word embedding models and Conditional Random Fields (CRF) for prediction. Li et al. [46] developed NLP methods to extract medication information (e.g., drug name, indication, contraindication) and adverse events from the FDA drug labels. Duke and Friedlin [47] applied MetaMap to identify ADEs from structured product labels.

Related works on corpora for clinical NLP research include the GENIA corpus [48] and the TREC Genomics [49]. Shared tasks such as BioNLP [34] and BioCreAtivE [35] have been widely used to train NLP applications. Other annotated corpora include the disease corpus [50], the BioScope corpus [51] and the MEDLINE abstract corpus from Gurulingappa et al. [52]. The corpus closest to ours is the i2b2 2009 corpus by Uzun et al. [53] which provides annotations for medication and related named entities. However, our work extends the annotation schema used in the i2b2 corpus and provides a common dataset for medication and adverse drug events. Another similar work is by Henriksson et al. [54], in which they

annotated a dataset focused towards ADE extraction. In contrast to their work, MADE corpus also provides annotation for medication details like dosage frequency etc. that are extremely relevant for PhV studies.

3 The MADE Corpus

The MADE corpus is composed of 1,089 fully de-identified longitudinal EHR notes from 21 randomly selected cancer patients at the University of Massachusetts Memorial Medical Center. Therefore, the notes include diverse note types such as discharge summaries, consultation reports, and other clinic notes.

We used an iterative process throughout the annotation, going back and forth between document annotations and establishing annotation guidelines. In this process, we created a comprehensive annotation guideline¹, which addresses various aspects on how to handle language variations and ambiguities in clinical narratives related to this annotation task. The guideline has adapted and substantially extended the 2009 i2b2 shared task of Medication Challenge annotation guideline [53]. The MADE annotation guideline is designed with a focus on extracting adverse drug events, and other relevant clinical information. It defines nine named entity types and seven relation types. The relation types define relationships between pairs of annotated named entities. A succinct overview of the annotation categories is provided below. The entities and relation types are described in detail in the text of the annotation guideline.

3.1 Named Entity Types

The named entity types can be broadly defined as either events or attributes. Events are annotations that denote a change in patient's medical status. This includes the prescription of a medication and identification of a symptom or diagnosis. Events have attributes, including severity and information related to medications (e.g., dosage). The occurrence of each named entity type is provided in Table 2. As evident from the table, there is a large label imbalance in the data, which makes it a challenging task to develop an NLP system to detect those entity types.

Based on their context, the named entity annotations can be clustered into those related to Sign Symptom or Disease (SSD) mentions, and those related to Medication (Drugname) mentions. The two categories are described below.

3.1.1 Sign, Symptom or Disease (SSD)—Annotations in the SSD group define events and properties relevant to SSD mentions. The relevant named entity types are adverse drug events, indication, other SSD and severity. Adverse drug events, indication, and other SSD are event annotations.

Adverse Drug Event (ADE): ADEs are a type of SSDs. They are adverse events caused by a Drugname. An ADE annotation requires a direct linguistic cue that links the adverse effect to a Drugname. E.g. "Patient had anaphylaxis after getting penicillin."

¹The complete annotation guideline and dataset is available at bio-nlp.org/dataset/made1

Indication: An indication is annotated if it is explicitly linked to a medication. E.x. “The patient was troubled with mouth sores and is being treated with Actiq.”

Other SSD: Any SSD event that is not annotated as an Indication or ADE is categorized as Other SSD. In our EHRs, Other SSDs frequently occur in the history section of notes. E.g. “headache in the back of the head”.

Severity: These annotations are attributes of SSDs, which indicate the severity (E.g. “acute”, “mild”, and “severe”) of a particular sign, symptom or disease.

3.1.2 Medication (Drugname)—Medication includes drugname and its attributes.

Drugname: Drugname annotation includes descriptions that denote any medication, procedure or therapy prescription. E.g. “Warfarin”, “Propofol”, “Chemotherapy” etc.

Duration: Duration is the time range for the administration of the Drugname, as explicitly described in the notes. E.g. “two weeks” “15 hours”.

Dosage: Dosage is the amount of drug in a unit dose. It is a numerical value and is an attribute of Drugname entity. E.g. “2 tablets”, “4 ml/hr.”.

Frequency: Frequency is the rate of administration of the drug and is an attribute of Drugname. E.g. “every hour”, “t.i.d.”, “q.i.d.”.

Route: Route is the path through which a drug is taken into the body. It is an attribute of Drugname entity. E.g. “orally”, “central line”.

3.2 Relation types

Table 3 shows the 7 relation types and their frequencies in the MADE corpus. A relation type is defined as a relation between two different named entity types. A brief description of each relation along with the relevant named entities is provided below.

Drugname Attribute Relations—MADE corpus contains 4 different relation types which describe a relation between the Drugname entity and its various attributes. These relations are:

- Drugname - Dosage
- Drugname - Route
- Drugname – Frequency
- Drugname – Duration

The attributes (Dosage, Route, Frequency, Duration) are properties of the drugname Entity.

SSD - Severity: Severity is an attribute to a Sign, Symptom or disease (ADE, Indication, Other SSD). It is typically a modifier (e.g., “mild”) for an annotated entity (for example “fever”).

ADE - Drugname: Adverse Drug Event is an adverse effect of the prescription of the Drugname Entity.

Indication - Drugname: The Drugname entity has been prescribed as a direct treatment for the Indication Entity.

In MADE corpus, the relations between the named entities can occur within a sentence or across multiple sentences in a note. Table 3 provides the relation length in characters. The ADE - Drugname and Indication - Drugname relations have heavy long tails, indicating that in several instances they connect named entities that are several sentences apart. We discuss the implications of this trend in Section 4.3.

3.3 Annotators

The annotation process involved multiple annotators, including physicians, biologists, linguists, and biomedical database curators. Annotators were used both in document annotation as well as the development of annotation guidelines. The following process was used to annotate each file. The first annotator individually labeled the span and type of named entities and relations. A second annotator then reviewed the annotations and modified them to produce the final version. This annotation process was used in order to reduce the annotation cost of each document while ensuring high annotation quality.

Since the annotations provided by the two annotators in this process are not independent, they cannot be used to obtain estimates of inter-annotator agreements (IAA). In order to get a fair IAA estimate, we performed a smaller study where five annotators independently annotated three documents from our corpus. We used the Fleiss' Kappa (k) [55] measure of inter-annotator agreement. The k for the named entity annotation and relation annotation agreement are 0.628 and 0.424 respectively. The relation k measures the agreement in both named entity and relation prediction. The added complexity of combined named entity and relation annotation may explain its comparatively lower annotation agreement value.

However, both values fall in the fair to significant agreement range [56], suggesting that our annotations are reliable for evaluating information extraction systems.

3.4 De-identification

The EHR data de-identification is done by the Safe Harbor methods defined in 45 CFR 164.514b(2) by the US Department of Health and Human Services. Firstly, the EHR data was processed by a publicly available de-identifier [57] so that the 18 types of Safe Harbor identifiers would be automatically annotated. Secondly, each clinical note was manually reviewed to ensure all identifiers were marked fully and correctly during the annotation process. All the marked identifiers were finally removed before releasing the data to the participating teams of MADE challenge.

3.5 Evaluation Script

To standardize the evaluation of NER and RI, we have developed an evaluation script². Our evaluation script uses bioc³ format, a simple format developed by the research community to share text data and annotations.

We used exact phrase-based evaluation, i.e., a named entity is correct only when the predicted span and entity type exactly matches the references annotation. This is important as partial matching (e.g., “infarction”) may be semantically different from the exact matching (e.g., “myocardial infarction”). For relations, a predicted relation between two entity types is regarded as correct only if both the relation type is correct and the prediction of all relevant named entities are correct. We used F-score for our system evaluation because it combines precision and recall, both of which are important metrics in evaluation of information extraction systems. A micro-averaged f-score was used to get an aggregate f-score over all classes. We strictly follow the micro-average implementation used by scikit-learn⁴. We also report both micro-averaged precision and recall score for the systems in the interest of interpretability.

Our evaluation script also provides an approximate metric which uses word-based evaluation, i.e., a named entity is correct if one or more words match. However approximate match is not used for evaluation in the MADE challenge. During the course of the challenge, we found instances of inconsistent inclusion or exclusion of the period for a named entity. Therefore, our evaluation script ignores span errors of one trailing character length in order to account for such inconsistencies. Details regarding the annotation inconsistencies can be found in Section 5.

3.6 Test and Train Data

In total, 213 notes from our MADE corpus were selected for the testing split and the remaining 876 notes formed the training split of MADE challenge. To minimize the potential of over-fitting and maximize the evaluation quality, we used two approaches to select the test set. We first selected three patients (out of 21) from the MADE cohort and included all their EHR notes (a total of 153) in the test set. We then selected 0 to 4 notes from the remaining 18 patients, in order to add an additional 60 notes for the test set.

4 The MADE Challenge

The MADE challenge invited participants to submit systems for three shared tasks. The MADE corpus training data was released in November 2017, four months before the final test run. Submissions were evaluated using the criterion described in Section 3.5. We designed two different runs, namely standard and extended. Submissions in the standard runs were limited in the type of external tools they could use, as such, it provides a fair evaluation of the NLP models. For this run, teams had no restriction in using open-NLP systems. They could use outputs of open NLP tools like Stanford NLP, NLTK and the UMLS tools for feature engineering. However, they were not allowed to use custom clinical NLP software, other EHR datasets or NLP tools trained on other EHR datasets. They were also not allowed to use proprietary or in-house NLP software. Systems not adhering to these constraints were categorized as extended runs. We allowed two submissions for each run from each participating team. However, we only considered standard runs in the final

²Evaluation script is include with the MADE data release

³<http://bioc.sourceforge.net>, <https://github.com/yfpeng/bioc>

⁴http://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

evaluation. Please refer to the relevant articles for an evaluation and analysis of their extended runs.

The three shared tasks are designed to evaluate submissions with an overall goal of identifying adverse drug events and other relevant entities and relations from EHR notes.

Task 1: Named Entity Recognition(NER)

This task requires extraction of both EHR named entities spans and their types from EHR notes. The named entity types are described in Section 3.1. Input is an unlabeled, raw EHR note. Output is a bioC file containing the entity span and type. The evaluation for this task uses f-score evaluation metric based on exact phrase matches, described in Section 3.5.

Task 2: Relation Identification (RI)

This task requires classification of relation and its type between two provided named entities. Since the named entities are provided as input, this task does not require detection of named entities span or types. The relation types are described in Section 3.2. The input is the unlabeled EHR notes and a bioC file containing the list of present named entities. The output is a bioC file containing the relationships between the provided list of named entities if any. The evaluation for this task uses f-score evaluation metric described in Section 3.5.

Task 3: Joint Relation Extraction (NER-RI)

This task requires prediction of both the named entities and their relations. Therefore, the systems submitted to this task have to jointly do both Named Entity Recognition and Relation Identification. Input is an unlabeled EHR document. Submissions are expected to correctly extract the named entities, predict the entity type and predict their relations. Output is a bioC file containing the named entity and relations. The evaluation for this task uses f-score evaluation metric based on exact phrase matches, as described in Section 3.5.

4.1 Submissions

The workshop submissions include 41 runs from 11 teams. The largest participation was in the first (NER) task and the smallest in the third (NER-RI). We show the exact F-score, Precision and Recall for all teams in Tables 4 (NER task), 5 (RI task), and 6 (NER-RI task). Table 7 provides a tabular view of the features and methods used. The NER task (Task 1) has a best F-score of 0.829. RI and NER-RI tasks have best F-scores of 0.8684 and 0.6170 respectively. We also evaluated the extended runs, although they were not used for MADE task rankings. No extended run performed better than the top system in any of the three tasks. The label-wise recall, precision and f-scores of the best submission system in all three tasks are provided in Tables 8, 9 and 10. These tables also provide the score for an ensemble prediction system composed of the top three runs in each task. More details about the ensemble system are provided in Section 4.3.

A brief overview of the methods used is provided in the following section. Please refer to the relevant team papers for further details about their methodologies.

4.2 Methods

As shown in table 7, while the teams developed a variety of different sequence labeling and machine learning models, Long Short-Term Memory and Conditional Random Fields were the most widely used models for named entity recognition. The relation identification task shows a variety of models such as SVM, Random Forest etc. A brief overview of the methods used in named entity recognition and relation classification are provided below.

Named Entity Recognition—The task of named entity recognition can be posed as a sequence labeling problem, where a sentence can be treated as a sequence of tokens. The task is then reduced to the problem of labeling each token with the named entity tag or an “Outside” (no named entity) tag. Commonly used algorithms for sequence labeling are Markov models (Hidden Markov Models, Conditional Random Fields), neural network models (Convolutional Neural Network, Recurrent Neural Network) or a combination of both.

Linear-chain Conditional Random Fields(CRF) [66] and related models (MEMM [67], HMM [68]) belong to a class of methods in machine learning based on Markov Models. CRF, which is a widely used method in sequence labeling, maximizes the joint probability of the label sequence conditioned on the input sentence.

Recurrent Neural Networks (RNN) such as Long Short-Term Memory [69] or Gated Recurrent Units (GRU) [70] are neural networks with recurrent connections that are designed to process sequential data. They have been shown to be useful in several natural language processing tasks such as NER [71], language modeling [72] and part of speech [73]. CRFs and RNNs were the two main methods used in MADE for the NER task. These methods take bag-of-word and other relevant features as input and produce the labels as outputs. Several teams experimented with character embeddings and other sub-word representations like suffix and prefix embeddings. Features such as part of speech, surface features (case-based features) were also used with both RNN and CRF based models. All teams use pre-trained word embedding to either pre-initialize their neural models or as features for CRF training.

Relation Classification—RI and NER-RI tasks require classification of the named entity pairs into several relation classes. The absence of a relation between the named entity pairs can be treated as another class in a multi-class classification scheme. Some submissions divide the classification into two separate sequential classification steps. The first classification task is to predict whether a relationship exists between the two named entity pairs. The second classification task predicts the type of that relation.

The classification methods range from neural network based methods such as a bidirectional LSTM with attention layer [59] to random forests [64, 65] and Support Vector Machines [61]. Neural network based methods use a final soft-max layer and cross entropy loss for training the relation classifier. Support Vector Machine [74] is a statistical machine learning technique which uses maximum margin loss to train the classifier. Random Forests [75] are a class of ensemble methods. They use the combined score from a collection of decision trees in order to produce the class prediction.

4.3 Analysis

The micro-average f-score for task 3 is significantly lower than tasks 1 and 2. This is to be expected since prediction in task 3 compounds the errors in both the NER and RI steps. For the real-world application of extracting Drugname and related ADE, the f-score needs to be further improved from its current best of 0.4272 in the NER-RI task (Table 10). A major factor behind the low score of the ADE-Drugname relation type is the low NER f-score of ADE. However, the prediction of this relation itself is also a challenging prediction problem as evidenced by the f-score of 0.72 in Task 2. This may be due to the fact that the text span between two entities in this relation could be large (Table 3). Similar arguments can also be made about the relation Indication-Drugname, which is another important relation type for downstream applications like drug-efficacy studies.

We ran paired sample t-test to evaluate the statistical significance of the differences between the top three models in each task. Paired t-test evaluates the difference between 2 related variables. The differences were considered significant if the p-value is less than 0.05. The samples used in our test were obtained by using file level micro-average f-score. For the first task, we found no statistically significant difference between the first [58] and second [59] system. However, the third system [60] was statistically significantly different from both Wunnava et al. [58] and Dandala et al. [59]. For task2, all differences between the top three teams were statistically significant. For task3, the third system [62] was statistically significantly different than the first [59]. All other differences among the top 3 teams in this task were statistically insignificant.

We built an ensemble system using the submitted runs. The ensemble output for each task is generated using a simple majority vote scheme. A prediction (named entity or a relation instance), is used by the ensemble if a majority of submissions agree on it. For shared task 1, the entire named entity phrase along with its type is taken as one prediction instance. For tasks 2 and 3, the complete relation prediction (relation type along with its constituent named entity predictions) constitutes one instance. The ensemble f-score is shown in table 11. Each ensemble is created by choosing one best standard run from each team. We also use an ensemble composed of one standard run each from the top three teams for each task. The ensembles show significant performance gains in task 2 and 3 when compared to the best individual system in each shared task category. Even in task 1, the performance increase in F-score is around 0.02. This indicates that the top systems in the MADE challenge do not all learn the same pattern from the dataset. Instead, there is variability in the information they learn.

The NER and NER-RI tasks are interesting, not only from a research perspective but they also have applications as steps in practical information extraction pipelines. It is non-trivial to accurately estimate an F-score threshold for a good real-world performance. Therefore, we cannot calibrate an F-score of 0.8 in the context of its real-world performance. However, in our experience, a precision score of 0.83 suggests that the system can extract reasonably accurate and useful data from unstructured text. Therefore, we believe that these models should be good enough for large-scale statistical studies where count-based thresholds can be used to reduce the noise in the extracted data. However, applications which require patient-specific information may need NER systems with higher recall and precision. For

instance, systems which use statistical methods to predict the outcome on a patient level may be very sensitive to the noise introduced by MADE NER systems.

As mentioned previously, the NER-RI performance is markedly lower than the NER performance of the submitted systems. The precision of NER-RI systems is significantly improved by building an ensemble as shown in Table 11. However, a relatively low F-score of around 0.6 suggests that the current NER-RI systems need to be further improved in order to be useful in real-world applications. Future steps in improving these models can focus on 1) improving the machine learning models, 2) annotation efforts in order to build larger labeled corpora, or 3) designing machine learning techniques that use external knowledge and unlabeled text.

5 Corpus Errors

The annotations in the corpus contain a few inconsistencies and errors. Some of these errors were observed while testing the evaluation script on the data, while several of them were reported by the participating teams. The errors fall into two categories, inconsistency in annotations and overlapping annotations.

The inconsistency in annotations is due to inconsistent annotations of the period character in named entity annotations. As an example, for the phrase “q.i.d.”, annotations may sometimes miss the last period character “q.i.d”. This inconsistency is only exhibited with the period character and only when it is a trailing period. In order to account for this inconsistency, the evaluation script ignores trailing mistakes of one-character length for all tasks.

Errors due to overlapping annotations occur when spans of two named entities overlap. The two overlapping entities can be of the same or different types. A common error in this category is overlapping annotations in the same type. For example, both the phrase “Vitamin D” and the overlapping sub-phrase “Vitamin” are annotated as separate Drugname annotations. Another common error in this category is double annotation on the same named entity. For example, the same phrase span “nausea” is annotated twice as ADE and Other SSD. Since it is not trivial to disambiguate the correct annotation for these errors, the evaluation script addresses this issue by treating all reference annotations as correct. This essentially means that the evaluation script slightly underestimates the true scores. However, since these errors are exhibited only in around 130 named entity annotations (from over seventy thousand NE annotations in MADE) the evaluation script score is still able to accurately access the performance of submitted systems.

6 Conclusion

We have created an expert-curated corpus composed of longitudinal EHR notes from patients with cancer. The MADE cohort was annotated with medication and adverse drug event related information. We released this cohort to the research community and used it as the benchmark to evaluate state of the art NLP models. MADE results show that recent progress in NLP has led to remarkable improvements in NER and RI tasks for the clinical domain. However, as demonstrated by the joint NER-RI task, there is still room for

improvement. We invite future research efforts in order to improve the state of the art on these benchmarks.

Acknowledgment

We are extremely thankful to the MADE 1.0 Annotation team: Elaine Freund, Heather Keating, Nadya Frid, Edgard Granillo, Raelene Goodwin, Brian Corner, Zuofeng Li, Rashmi Prasad, Balaji Ramesh, Victoria Wang and Steven Belknap for their contributions in the MADE project. They were an essential part of the data curation, annotation, and research process for MADE 1.0. They are also authors of the annotation guideline used throughout the development of this corpus.

Funding

Research reported in this publication was supported by the National Heart, Lung, and Blood Institute (NHLBI) of the National Institutes of Health under Award Number R01HL125089.

References

- [1]. Donaldson Molla S, Corrigan Janet M, Kohn Linda T, et al. To err is human: building a safer health system, volume 6. National Academies Press, 2000.
- [2]. Bates David W, Cullen David J, Laird Nan, Petersen Laura A, Small Stephen D, Servi Deborah, Laffel Glenn, Sweitzer Bobbie J, Shea Brian F, Hallisey Robert, et al. Incidence of adverse drug events and potential adverse drug events: implications for prevention. *Jama*, 274(1):29–34, 1995. [PubMed: 7791255]
- [3]. Lazarou Jason, Pomeranz Bruce H, and Corey Paul N. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *Jama*, 279 (15):1200–1205, 1998. [PubMed: 9555760]
- [4]. Bate David W, Spell Nathan, Cullen David J, Burdick Elisabeth, Laird Nan, Petersen Laura A, Small Stephen D, Sweitzer Bobbie J, and Leape Lucian L. The costs of adverse drug events in hospitalized patients. *Jama*, 277(4):307–311, 1997. [PubMed: 9002493]
- [5]. Nebeker Jonathan R, Hoffman Jennifer M, Weir Charlene R, Bennett Charles L, and Hurdle John F. High rates of adverse drug events in a highly computerized hospital. *Archives of internal medicine*, 165(10):1111–1116, 2005. [PubMed: 15911723]
- [6]. Gurwitz Jerry H, Field Terry S, Harrold Leslie R, Rothschild Jeffrey, Debellis Kristin, Seger Andrew C, Cadoret Cynthia, Fish Leslie S, Garber Lawrence, Kelleher Michael, et al. Incidence and preventability of adverse drug events among older persons in the ambulatory setting. *Jama*, 289(9):1107–1116, 2003. [PubMed: 12622580]
- [7]. Johnson Jeffery and Booman Lyle. Drug-related morbidity and mortality. *Journal of Managed Care Pharmacy*, 2(1):39–47, 1996.
- [8]. Haas Jennifer S, Iyer Aarthi, Orav E John, Schiff Gordon D, and Bates David W. Participation in an ambulatory e-pharmacovigilance system. *Pharmacoepidemiology and drug safety*, 19(9):961–969, 2010. [PubMed: 20623512]
- [9]. Frank Cassie, Himmelstein David U, Woolhandler Steffie, Bor David H, Wolfe Sidney M, Heymann Orlaith, Zallman Leah, and Lasser Karen E. Era of faster fda drug approval has also seen increased black-box warnings and market withdrawals. *Health affairs*, 33(8):1453–1459, 2014. [PubMed: 25092848]
- [10]. WHO. WHO | Pharmacovigilance., 2017 http://www.who.int/medicines/areas/quality_safety/safety_efficacy/pharmvigi/en/.
- [11]. Edlavitch Stanley A. Adverse drug event reporting: improving the low us reporting rates. *Archives of internal medicine*, 148(7):1499–1503, 1988. [PubMed: 3382293]
- [12]. Hasford J, Goettler M, Munter K-H, and Müller-Oerlinghausen B. Physicians' knowledge and attitudes regarding the spontaneous reporting system for adverse drug reactions. *Journal of clinical epidemiology*, 55(9):945–950, 2002. [PubMed: 12393084]

- [13]. Begaud B, Moride Y, Tubert-Bitter P, Chaslerie A, and Haramburu F. False-positives in spontaneous reporting: should we worry about them? *British journal of clinical pharmacology*, 38(5):401–404, 1994. [PubMed: 7893579]
- [14]. Xu Rong and Wang QuanQiu. Comparing a knowledge-driven approach to a super-vised machine learning approach in large-scale extraction of drug-side effect relation-ships from free-text biomedical literature. In *BMC bioinformatics*, volume 16, page S6. BioMed Central, 2015.
- [15]. Butt Tehreem F, Cox Anthony R, Oyeboode Jan R, and Ferner Robin E. Internet accounts of serious adverse drug reactions. *Drug safety*, 35(12):1159–1170, 2012. [PubMed: 23058037]
- [16]. Rossi Allen C, Knapp Deanne E, Anello Charles, O'Neill Robert T, Graham Cheryl F, Mendelis Peter S, and Stanley George R. Discovery of adverse drug reactions: a comparison of selected phase iv studies with spontaneous reporting methods. *Jama*, 249(16):2226–2228, 1983. [PubMed: 6834622]
- [17]. Lardon Jérémy, Abdellaoui Redhouane, Bellet Florelle, Asfari HadyI, Souvi-gnet Julien, Texier Nathalie, Jaulent Marie-Christine, Beyens Marie-Noëlle, Burgun Anita, and Bousquet Cédric. Adverse drug reaction identification and extraction in social media: a scoping review. *Journal of medical Internet research*, 17(7), 2015.
- [18]. Maureen A Smythe, Fanikos John, Gulseth Michael P, Wittkowsky Ann K, Spinler Sarah A, Dager William E, and Nutescu Edith A. Rivaroxaban: practical consider-ations for ensuring safety and efficacy. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 33(11):1223–1245, 2013.
- [19]. McGraw Deven, Rosati Kristen, and Evans Barbara. A policy framework for public health uses of electronic health data. *Pharmacoepidemiology and drug safety*, 21(S1): 18–22, 2012. [PubMed: 22262589]
- [20]. Yih W Katherine, Lieu Tracy A, Kulldorff Martin, Martin David, McMahonill-Walraven Cheryl N, Platt Richard, Selvam Nandini, Selvan Mano, Lee Grace M, and Nguyen Michael. Intussusception risk after rotavirus vaccination in us infants. *New England Journal of Medicine*, 370(6):503–512, 2014. [PubMed: 24422676]
- [21]. Peissig Peggy L, Costa Vitor Santos, Caldwell Michael D, Rottscheit Carla, Berg Richard L, Mendonca Eneida A, and Page David. Relational machine learning for elec-tronic health record-driven phenotyping. *Journal of biomedical informatics*, 52:260–270, 2014. [PubMed: 25048351]
- [22]. Wu Jionglin, Roy Jason, and Stewart Walter F. Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches. *Medical care*, pages S106–S113, 2010.
- [23]. Jha Ashish K, Kuperman Gilad J, Teich Jonathan M, Leape Lucian, Shea Brian, Rittenberg Eve, Burdick Elisabeth, Seger Diane Lew, Vliet Martha Vander, and Bates David W. Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report. *Journal of the American Medical Informatics Association*, 5(3):305–314, 1998. [PubMed: 9609500]
- [24]. Skentzos Stephen, Shubina Maria, Plutzky Jorge, and Turchin Alexander. Structured vs. unstructured: factors affecting adverse drug reaction documentation in an emr repository. In *AMIA Annual Symposium Proceedings*, volume 2011, page 1270 American Medical Informatics Association, 2011.
- [25]. Schulman S, Kearon C, Subcommittee on Control of Anticoagulation of the Scientific, Standardization Committee of the International Society on Thrombosis, and Haemostasis. Definition of major bleeding in clinical investigations of antihemostatic medicinal products in non-surgical patients. *Journal of thrombosis and haemostasis*, 3(4):692–694, 2005. [PubMed: 15842354]
- [26]. Murtaugh Maureen A, Gibson Bryan Smith, Redd Doug, and Zeng-Treitler Qing. Regular expression-based learning to extract bodyweight values from clinical notes. *Journal of biomedical informatics*, 54:186–190, 2015. [PubMed: 25746391]
- [27]. Classen David C, Pestotnik SL, Evans RS, and Burke JP. Computerized surveillance of adverse drug events in hospital patients. *BMJ Quality & Safety*, 14(3):221–226, 2005.
- [28]. Aronson Alan R. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17 American Medical Informatics Association, 2001.

- [29]. Xu Hua, Stenner Shane P, Doan Son, Johnson Kevin B, Waitman Lemuel R, and Denny Joshua C. Medex: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24, 2010. [PubMed: 20064797]
- [30]. Friedman Carol, Kra Pauline, Yu Hong, Krauthammer Michael, and Rzhetsky Andrey. Genies: a natural-language processing system for the extraction of molecular pathways from journal articles. In *ISMB (supplement of bioinformatics)*, pages 74–82, 2001.
- [31]. Hahn Udo, Romacker Martin, and Schulz Stefan. Creating knowledge repositories from biomedical reports: the medsyndicate text mining system In *Biocomputing 2002*, pages 338–349. World Scientific, 2001.
- [32]. Yu Hong and Lee Minsuk. Accessing bioscience images from abstract sentences. *Bioinformatics*, 22(14):e547–e556, 2006.
- [33]. Yu Hong. Towards answering biological questions with experimental evidence: automatically identifying text that summarize image content in full-text articles. In *AMIA Annual Symposium Proceedings*, volume 2006, page 834 American Medical Informatics Association, 2006.
- [34]. Kim Jin-Dong, Ohta Tomoko, Pyysalo Sampo, Kano Yoshinobu, and Tsujii Jun'ichi. Overview of bionlp'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 1–9 Association for Computational Linguistics, 2009.
- [35]. Hirschman Lynette, Yeh Alexander, Blaschke Christian, and Valencia Alfonso. Overview of biocreative: critical assessment of information extraction for biology, 2005.
- [36]. Li Zuofeng, Cao Yonggang, Antieau Lamont, Agarwal Shashank, Zhang Qing, and Yu Hong. Extracting medication information from patient discharge summaries. In *Proceedings of the third i2b2 workshop on challenges in natural language processing for clinical data*, 2009.
- [37]. Pradhan Sameer, Elhadad Noémie, Brett R South David Martinez, Christensen Lee, Vogel Amy, Suominen Hanna, Chapman Wendy W, and Savova Guergana. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *Journal of the American Medical Informatics Association*, 22(1):143–154, 2014. [PubMed: 25147248]
- [38]. Li Qi, Melton Kristin, Lingren Todd, Eric S Kirkendall Eric Hall, Zhai Haijun, Ni Yizhao, Kaiser Megan, Stoutenborough Laura, and Solti Imre. Phenotyping for patient safety: algorithm development for electronic health record based automated adverse event and medical error detection in neonatal intensive care. *Journal of the American Medical Informatics Association*, 21(5):776–784, 2014. [PubMed: 24401171]
- [39]. Melton Genevieve B and Hripcsak George. Automated detection of adverse events using natural language processing of discharge summaries. *Journal of the American Medical Informatics Association*, 12(4):448–457, 2005. [PubMed: 15802475]
- [40]. Humphreys Betsy L, Lindberg Donald AB, Schoolman Harold M, and Barnett G Octo. The unified medical language system: an informatics research collaboration. *Journal of the American Medical Informatics Association*, 5(1):1–11, 1998. [PubMed: 9452981]
- [41]. Bodenreider Olivier. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004. [PubMed: 14681409]
- [42]. Rochefort Christian M, Verma Aman D, Eguale Tewodros, Lee Todd C, and Buckeridge David L. A novel method of adverse event detection can accurately identify venous thromboembolisms (vtes) from narrative electronic health record data. *Journal of the American Medical Informatics Association*, 22(1):155–165, 2014. [PubMed: 25332356]
- [43]. Haerian Krystl, Salmasian Hojjat, and Friedman Carol. Methods for identifying suicide or suicidal ideation in ehrrs. In *AMIA Annual Symposium Proceedings*, volume 2012, page 1244 American Medical Informatics Association, 2012.
- [44]. Wang Sheng, Li Yanen, Ferguson Duncan, and Zhai Chengxiang. Sideeffectptm: An unsupervised topic model to mine adverse drug reactions from health forums. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 321–330 ACM, 2014.
- [45]. Nikfarjam Azadeh, Sarker Abeed, Karen O'Connor Rachel Ginn, and Gonzalez Graciela. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence

- labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681, 2015. [PubMed: 25755127]
- [46]. Li Qi, Deleger Louise, Lingren Todd, Zhai Haijun, Kaiser Megan, Stoutenborough Laura, Jegga Anil G, Cohen Kevin Bretonnel, and Solti Imre. Mining fda drug labels for medical conditions. *BMC medical informatics and decision making*, 13(1):53, 2013. [PubMed: 23617267]
- [47]. Duke Jon D and Friedlin Jeff. Adessa: a real-time decision support service for delivery of semantically coded adverse drug event data. In *AMIA Annual Symposium Proceedings*, volume 2010, page 177 American Medical Informatics Association, 2010.
- [48]. Kim J-D, Ohta Tomoko, Tateisi Yuka, and Jun'ichi Tsujii. Genia corpus-a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182, 2003. [PubMed: 12855455]
- [49]. Cohen Aaron M and Hersh William R. The trec 2004 genomics track categorization task: classifying full text biomedical documents. *Journal of Biomedical Discovery and Collaboration*, 1(1):4, 2006. [PubMed: 16722582]
- [50]. Do an Rezarta Islamaj and Lu Zhiyong. An improved corpus of disease mentions in pubmed citations. In *Proceedings of the 2012 workshop on biomedical natural language processing*, pages 91–99 Association for Computational Linguistics, 2012.
- [51]. Vincze Veronika, Szarvas György, Farkas Richárd, Móra György, and Csirik János. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(11):S9, 2008.
- [52]. Gurulingappa Harsha, Rajput Abdul Mateen, Roberts Angus, Fluck Julianne, Hofmann-Apitius Martin, and Toldo Luca. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892, 2012. [PubMed: 22554702]
- [53]. Uzuner Özlem, Solti Imre, and Cadag Eithon. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5): 514–518, 2010. [PubMed: 20819854]
- [54]. Henriksson Aron, Kvist Maria, Dalianis Hercules, and Duneld Martin. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. *Journal of biomedical informatics*, 57:333–349, 2015. [PubMed: 26291578]
- [55]. Fleiss Joseph L. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.
- [56]. Landis J Richard and Koch Gary G. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.
- [57]. Liu Zengjian, Chen Yangxin, Tang Buzhou, Wang Xiaolong, Chen Qingcai, Li Haodi, Wang Jingfeng, Deng Qiwen, and Zhu Suisong. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. *Journal of biomedical informatics*, 58:S47–S52. doi: 10.1016/j.jbi.2015.06.009.
- [58]. Wunnavu Susmitha, Qin Xiao, Kakar Tabassum, Rundensteiner Elke A., and Kong Xiang-nan. Bidirectional lstmcrf for adverse drug event tagging in electronic health records. In Liu Feifan, Jagannatha Abhyuday, and Yu Hong, editors, *Proceedings of the 1st International Workshop on Medication and Adverse Drug Event Detection*, volume 90 of *Proceedings of Machine Learning Research*, pages 48–56 PMLR, 04 May 2018 URL <http://proceedings.mlr.press/v90/wunnavu18a.html>.
- [59]. Dandala B, Joopudi V, Devarakonda M. Adverse Drug Events Detection in Clinical Notes by Jointly Modeling Entities and Relations using Neural Networks. *Drug Saf* 2018.doi [TO BE ADDED DURING PRODUCTION]
- [60]. Yang X, Bian J, Gong Y, Hogan WR, Wu Y. MADEx: A System for Detecting Medications, Adverse Drug Events, and their Relations from Clinical Notes. *Drug Saf* 2018.doi [TO BE ADDED DURING PRODUCTION]
- [61]. Xu Dongfang, Yadav Vikas, and Bethard Steven. Uarizona at the made1.0 nlp challenge. In Liu Feifan, Jagannatha Abhyuday, and Yu Hong, editors, *Proceedings of the 1st International Workshop on Medication and Adverse Drug Event Detection*, volume 90 of *Proceedings of Machine Learning Research*, pages 57–65 PMLR, 04 May 2018 URL <http://proceedings.mlr.press/v90/xu18a.html>.

- [62]. Chapman AB, Peterson KS, Alba PR, DuVall SL, Patterson OV. Detecting adverse drug events with rapidly trained classification models. *Drug Saf* 2018.doi [TO BE ADDED DURING PRODUCTION]
- [63]. Ngo Duy-Hoa, Metke-Jimenez Alejandro, and Nguyen Anthony. Knowledge-based feature engineering for detecting medication and adverse drug events from electronic health records. In Liu Feifan, Jagannatha Abhyuday, and Yu Hong, editors, *Proceedings of the 1st International Workshop on Medication and Adverse Drug Event Detection*, volume 90 of *Proceedings of Machine Learning Research*, pages 31–38 PMLR, 04 May 2018 URL <http://proceedings.mlr.press/v90/ngo18a.html>.
- [64]. Magge Arjun, Scotch Matthew, and Gonzalez-Hernandez Graciela. Clinical ner and relation extraction using bi-char-lstms and random forest classifiers. In Liu Feifan, Jagannatha Abhyuday, and Yu Hong, editors, *Proceedings of the 1st International Workshop on Medication and Adverse Drug Event Detection*, volume 90 of *Proceedings of Machine Learning Research*, pages 25–30 PMLR, 04 May 2018 URL <http://proceedings.mlr.press/v90/magge18a.html>.
- [65]. Florez Edson, Precioso Frederic, Riveill Michel, and Pighetti Romaric. Named entity recognition using neural networks for clinical notes. In Liu Feifan, Jagannatha Abhyuday, and Yu Hong, editors, *Proceedings of the 1st International Workshop on Medication and Adverse Drug Event Detection*, volume 90 of *Proceedings of Machine Learning Research*, pages 7–15 PMLR, 04 May 2018 URL <http://proceedings.mlr.press/v90/florez18a.html>.
- [66]. Lafferty John, McCallum Andrew, and Pereira Fernando CN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.
- [67]. McCallum Andrew, Freitag Dayne, and Pereira Fernando CN. Maximum entropy markov models for information extraction and segmentation. In *Icml*, volume 17, pages 591–598, 2000.
- [68]. Zhou GuoDong and Su Jian. Named entity recognition using an hmm-based chunk tagger. In *proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 473–480 Association for Computational Linguistics, 2002.
- [69]. Gers Felix A, Schmidhuber Jürgen, and Cummins Fred. Learning to forget: Continual prediction with lstm 1999.
- [70]. Chung Junyoung, Gulcehre Caglar, Cho Kyunghyun, and Bengio Yoshua. Gated feedback recurrent neural networks. In *International Conference on Machine Learning*, pages 2067–2075, 2015.
- [71]. Jagannatha Abhyuday N and Yu Hong. Bidirectional rnn for medical event detection in electronic health records. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2016, page 473 NIH Public Access, 2016.
- [72]. Sundermeyer Martin, Schlüter Ralf, and Ney Hermann. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*, 2012.
- [73]. Huang Zhiheng, Xu Wei, and Yu Kai. Bidirectional lstmcrf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.
- [74]. Cristianini Nello, Shawe-Taylor John, et al. *An introduction to support vector machines and other kernelbased learning methods*. Cambridge university press, 2000.
- [75]. Breiman Leo. Random forests. *Machine learning*, 45(1):5–32, 2001.

Key Points

- The MADE 1.0 corpus is composed of 1089 electronic health records with detailed named entity and relation annotations.
- We provide benchmark results and analysis on the MADE 1.0 corpus using system submissions in the MADE 1.0 challenge.
- MADE 1.0 results suggest that machine learning systems can be useful for automated extraction of adverse drug events and related entities from unstructured texts, however, there is still room for improvement.

Table 1:

The overall statistics for the MADE corpus

Description	Mean	Std (Max, Min)
Words / Note	948.2	484.9 (3804,76)
Named Entity Annotations / Note	72.6	52.6 (363,0)
Relation Annotations / Note	25.0	24.1 (181,0)
Notes/ Patient	51.8	40.3 (166,1)

Table 2:

Annotation counts, and word counts for each named entity type.

NE type	Number of Annotations	Total annotated words
<i>ADE</i>	1940	3255
<i>Indication</i>	3804	8240
<i>Other SSD</i>	39384	82956
<i>Severity</i>	3908	5069
<i>Drugname</i>	15902	19075
<i>Dosage</i>	5694	11820
<i>Duration</i>	898	1768
<i>Frequency</i>	4806	11400
<i>Route</i>	2667	2805

Table 3:

Annotation counts and relation length for each relation category. A relation is defined as a relation between two named entities. The relation length format is “mean $\pm$ std (max,min)”.

Relation type	Occurrences	Relation length
<i>ADE - Drugname</i>	2612	82 $\pm$ 187 (3662,1)
<i>SSD - Severity</i>	4035	4.7 $\pm$ 34.41 (1861,0)
<i>Drugname - Route</i>	3006	18 $\pm$ 25(224,1)
<i>Drugname - Dosage</i>	6043	11 $\pm$ 22(230,0)
<i>Drugname - Duration</i>	1053	20 $\pm$ 27(273,1)
<i>Drugname - Frequency</i>	5149	25 $\pm$ 30(295,1)
<i>Indication - Drugname</i>	5430	96 $\pm$ 164 (2742,1)

Table 4:

Performance metrics for the best runs by teams for Named Entity Recognition (NER) task (Shared Task 1).

Ranking	Team Names	Recall	Precision	F-score
1	WPI-Wunnavu[58]	0.8247	0.8333	0.8290
2	IBMResearch-dandala[59]	0.8243	0.8327	0.8285
3	UFL-gators [60]	0.8148	0.8318	0.8232
4	UArizonaSchool-Xu [61]	0.8042	0.8272	0.8156
5	UofUtah-Patterson[62]	0.7667	0.8280	0.7962
6	AEHRC-HoaNGO [63]	0.7463	0.8349	0.7881
7	ASU-BMI[64]	0.7074	0.8358	0.7663
8	MITRE-Tresner-Kirsch	0.7581	0.7443	0.7511
9	SUNY-Tao	0.6658	0.8054	0.7290
10	UCA-I3S-SPARK [65]	0.7198	0.6814	0.7001

Table 5:

Performance metrics for the best runs by teams for Relation Identification (RI) task (Shared Task 2).

Ranking	Team Names	Recall	Precision	F-score
1	UofUtah-Patterson [62]	0.8806	0.8565	0.8684
2	IBMResearch-dandala[59]	0.8736	0.8093	0.8402
3	UArizonaIschool-Xu[61]	0.8846	0.7849	0.8318
4	ASU-BMI [64]	0.7693	0.8680	0.8157
5	KazanFederalUniversity-Alimova	0.6225	0.3343	0.4350

Table 6:

Performance metrics for the best runs by teams for the joint Relation Identification (NER-RI) task (Shared Task 3).

Ranking	Team Names	Recall	Precision	F-score
1	IBMResearch-dandala [59]	0.6317	0.6029	0.6170
2	UArizonaSchool-Xu [61]	0.6005	0.5965	0.5985
3	UofUtah-Patterson [62]	0.5176	0.6918	0.5921
4	ASU-BMI[64]	0.4350	0.6431	0.5189

Table 7:

Architecture details shared by teams within the workshop proceedings. PWE and CE refer to “Pre-trained Word Embeddings” and “Character Embeddings” respectively. The “Features” column documents the additional features used by the submission for their NER system. “Relation Classifier” contains the classification methodology used by the submission for relation classification.

Team Names	LSTM	CRF	PWE	CE	Features	Relation Classifier
UCA-I3S-SPARK [65]	+	–	+	+	POS	–
UFL-gators [60]	+	–	+	–	–	–
UofUtah-Patterson [62]	–	+	+	–	POS, Surface	Random Forest
ASU-BMI [64]	+	+	+	+	Surface	Random Forest
IBMResearch-dandala [59]	+	+	+	+	POS	Attention Bi-LSTM
WPI-Wunnava [58]	+	+	+	+	–	–
UArizonaSchool-Xu [61]	+	+	+	+	Prefix, Suffix Embedding	SVM
AEHRC-HoaNGO [63]	–	+	+	–	Snomed-CT, POS, Dependency	–

Table 8:

Label-wise recall, precision and f-score values for the top submission and ensemble in Task 1. The ensemble takes the majority vote from the top 3 systems.

NEType	Best Submission			Top 3 Ensemble		
	Recall	Precision	F-score	Recall	Precision	F-score
ADE	0.5266	0.7229	0.6093	0.5058	0.7956	0.6184
Indication	0.5959	0.6467	0.6202	0.5833	0.7860	0.6696
Other SSD	0.8490	0.8294	0.8391	0.8547	0.8483	0.8515
Drugname	0.8743	0.9057	0.8897	0.8922	0.9319	0.9116
Duration	0.7443	0.6428	0.6898	0.6766	0.9000	0.7725
Frequency	0.7541	0.7611	0.7576	0.8376	0.9123	0.8734
Dosage	0.8888	0.8757	0.8822	0.8926	0.8915	0.8920
Severity	0.8014	0.8492	0.8246	0.8014	0.8577	0.8286
Route	0.9383	0.9125	0.9252	0.9203	0.9396	0.9298

Table 9:

Label-wise recall, precision and f-score values for the top submission and ensemble in Task 2. The ensemble takes the majority vote from the top 3 systems.

Relation Type	Best Submission			Top 3 Ensemble		
	Recall	Precision	F-score	Recall	Precision	F-score
ADE – Drugname	0.7377	0.7032	0.7200	0.7207	0.7047	0.7126
SSD – Severity	0.9677	0.9359	0.9516	0.9534	0.9334	0.9433
Drugname –Route	0.9120	0.9346	0.9232	0.9274	0.9483	0.9377
Drugname –Dosage	0.9676	0.9544	0.9610	0.9665	0.9479	0.9571
Drugname –Duration	0.9047	0.7732	0.8338	0.9523	0.6278	0.7567
Drugname –Route	0.9260	0.9428	0.9343	0.9369	0.9647	0.9506
Indication –Drugname	0.7671	0.7187	0.7421	0.8242	0.7680	0.7951

Table 10:

Label-wise recall, precision and f-score values for the top submission and ensemble in Task 3. The ensemble takes the majority vote from the top 3 systems.

Relation Type	Best Submission			Top 3 Ensemble		
	Recall	Precision	F-score	Recall	Precision	F-score
ADE – Drugname	0.4264	0.4280	0.4272	0.3264	0.6784	0.4407
SSD – Severity	0.5080	0.4694	0.4879	0.4740	0.5875	0.5247
Drugname –Route	0.7956	0.7801	0.7878	0.7868	0.8689	0.8258
Drugname –Dosage	0.7840	0.7324	0.7573	0.7655	0.8371	0.7997
Drugname –Duration	0.4557	0.4685	0.4620	0.4965	0.5572	0.5251
Drugname –Route	0.7424	0.7970	0.7687	0.7260	0.9059	0.8060
Indication –Drugname	0.5365	0.4630	0.4970	0.4223	0.6525	0.5128

Table 11:

Performance metrics calculated for ensemble of submissions as described in Section 4.3 The “Top 3 Team” shows the ensemble generated from submissions of top 3 teams. The “All Teams” ensemble uses submissions from all teams. For each team in both ensembles, only the best run is used. So, all ensembles in “Top 3 Teams” are ensembles of three different systems.

Shared Task	Top 3 teams			All Teams		
	Recall	Precision	F-score	Recall	Precision	F-score
Task 1:NER	0.8334	0.87297	0.8527	0.7847	0.9094	0.8425
Task 2:RI	0.8935	0.8625	0.8777	0.8782	0.7845	0.8287
Task 3:NER-RI	0.5841	0.7616	0.6612	0.4633	0.8580	0.6017