Active Deep Learning-Based Annotation of Electroencephalography Reports for Cohort Identification

Ramon Maldonado[1], Travis R. Goodwin[1], Sanda M. Harabagiu, PhD[1],
[1]University of Texas at Dallas, Richardson, TX, U.S.A.

Abstract

The annotation of a large corpus of Electroencephalography (EEG) reports is a crucial step in the development of an EEG-specific patient cohort retrieval system. The annotation of multiple types of EEG-specific medical concepts, along with their polarity and modality, is challenging, especially when automatically performed on Big Data. To address this challenge, we present a novel framework which combines the advantages of active and deep learning while producing annotations that capture a variety of attributes of medical concepts. Results obtained through our novel framework show great promise.

Introduction

Clinical electroencephalography (EEG) is the most important investigation in the diagnosis and management of epilepsies. In addition, it is used to evaluate other types of brain disorders[1], including encephalopathies, neurological infections, Creutzfeldt-Jacob disease and other prion disorders, and even in the progression of Alzheimer's disease. An EEG records the electrical activity along the scalp and measures spontaneous electrical activity of the brain. The signals measured along the scalp can be correlated with brain activity, which makes it a primary tool for diagnosis of brain-related illnesses[2]. But, as noted in [3], the EEG signal is complex, and thus its interpretation documented in EEG reports is producing inter-observer agreement in EEG interpretation known to be moderate. As more clinical EEG becomes available, the interpretation of EEG signals can be improved by providing neurologists with results of search for patients that exhibit similar EEG characteristics. Recently, Goodwin & Harabagiu (2016)[4] have described the MERCuRY (Multi-modal EncephalogRam patient Cohort discoveRY) system that uses deep learning to represent the EEG signal and operates on a multi-modal EEG index resulting from the automatic processing of both the EEG signal and the EEG reports that document and interpret them. The MERCuRY system allows neurologist to search a vast data archive of clinical electroencephalography (EEG) signals and EEG reports, enabling them to discover patient populations relevant to queries like *Q: Patients taking topiramate (Topomax) with a diagnosis of headache and EEGs demonstrating sharp waves, spikes or spike/polyspike and wave activity*

The discovery of relevant patient cohorts satisfying the characteristics expressed in queries such as *Q* relies on the ability of automatically and accurately recognizing both in the queries and throughout the EEG reports various medical concepts and their attributes. For example, a patient from this cohort could be identified if the following annotations indicating medical problems [PROB], treatments [TR], tests [TEST], EEG activities [ACT], and EEG events [EV] would be available in various sections of its EEG report:

> Example 1: *CLINICAL HISTORY: Recently [seizure]$_{PROB}$-free but with [episodes of light flashing in her peripheral vision]$_{PROB}$ followed by [blurry vision]$_{PROB}$ and [headaches] $_{PROB}$*
>
> *MEDICATIONS: [Topomax]$_{TR}$*
>
> *DESCRIPTION OF THE RECORD: There are also bursts of irregular, frontally predominant [sharply contoured delta activity]$_{ACT}$, some of which seem to have an underlying [spike complex]$_{ACT}$ from the left mid-temporal region.*

The relevance models implemented in the MERCuRY system would consider the annotations produced also on the query *Q* to discover the patients:

> *Q*$_{annotated}$: *Patients taking [topiramate]$_{MED}$ ([Topomax]$_{MED}$) with a diagnosis of [headache]$_{PROB}$ and [EEGs]$_{TEST}$ demonstrating [sharp waves]$_{ACT}$, [spikes]$_{ACT}$ or [spike/polyspike and wave activity]$_{ACT}$*

As big data for EEG becomes available, new deep learning techniques show promise for producing such annotations with high efficiency and accuracy. In this paper, we present a novel active learning framework that incorporates deep learning methods to annotate EEG-specific concepts and their attributes. An *EEG activity* is defined as "*an EEG wave or sequence of waves*", and an *EEG event* is defined as *"a stimulus that activates the EEG"* by the International Federation of Clinical Neurophysiology[5].

Background

Active learning (AL) has been proven to effectively reduce the amount of human annotation and validation when an efficient sampling mechanism is designed because it selects, for validation, those instances that impact the most the learning quality. In [6], an active-learning-based annotation that operates on MEDLINE abstracts was reported. Those annotations did not consider the modality or polarity of concepts, unlike the annotations produced by the Informatics for Integrating Biology and the Bedside (i2b2) 2010[7] and 2012[7] challenges. Hence, the active learning experience reported in [6] targets only the task of annotating medical concepts in biomedical text, ignoring their modality or polarity. In our novel framework, not only have we considered a large number of annotation tasks, but we perform them on a big corpus of EEG reports by taking advantage of new deep learning architectures which have recently produced very promising results[8]. Our deep learning architectures have allowed us to perform a significant number of annotation tasks concurrently without the burden of training a large number of classifiers.

Unlike previous annotation experiments on EHRs, we also tackled the case when a medical concept, in our case EEG activity, is not mentioned in a continuous span of text. For this purpose, we have defined the notion of *anchor* and *attributes* to be able to capture the characteristics of EEG activities. The anchor represents the morphology of an EEG activity, defined as the type or form of an EEG wave. EEG activities are always mentioned by referring to their morphology, thus this attribute "anchors" the concept mention. In contrast, the other EEG attributes are not always explicit, as EEG reports are written for an audience of neurologists, and are often implied. Hence, we needed to devise an annotation schema that captures the semantic richness of attributes of EEG activities. Moreover, many of the attributes, when expressed, may be mentioned at some distance from the anchor. E.g. in example 1, the anchor "sharply contoured delta activity" is far from the attribute expressed by "bursts" which represents the recurrence attribute of this EEG activity concept.

Data

In this work, we used a corpus of EEG reports available from the Temple University Hospital (TUH), comprising over 25,000 EEG reports from over 15,000 patients collected over 12 years. The EEG reports contain a great deal of medical knowledge, as they are designed to convey a written impression of the visual analysis of the EEG, along with an interpretation of its clinical significance. Following the American Clinical Neurophysiology Society Guidelines for writing EEG reports, the reports from the TUH EEG Corpus start with a clinical history of the patient, including information about the patient age, gender, conditions prevalent at the time of the recording (e.g., "after cardiac arrest") followed by a list of the medications (that might modify the EEG). Clearly, both of these initial sections depict the *clinical picture of the patient*, containing a wealth of medical concepts, including the medical problems (e.g. "cardiac arrest"), symptoms (e.g. "without a heart rate"), signs (e.g. "twitching") as well as significant medical events (e.g. "coded for 30 minutes in the emergency room") which are temporally grounded ("e.g. "30 minutes"). The following sections of the EEG report target mostly information related to the EEG techniques, interpretation and findings. The introduction section is the depiction of the techniques used for the EEG (e.g. "digital video EEG", "using standard 10-20 system of electrode placement with 1 channel of EKG"), as well as the patient's conditions prevalent at the time of the recording (e.g., fasting, sleep deprivation) and level of consciousness (e.g. "comatose"). The description section is the mandatory part of the EEG report, and it provides a complete and objective description of the EEG, noting all observed activity (e.g. "beta frequency activity"), patterns (e.g. "burst suppression pattern") and events ("very quick jerks of the head"). Many medical events mentioned in the description section of an EEG report are also grounded spatially ("e.g. "attenuated activity in the left hemisphere") as well as temporally (e.g. "beta frequency activity *followed by* some delta"). In addition, the EEG activities are characterized by a variety of attributes (e.g. strength "bursts of paroxysmal high amplitude activity"). The impression section states whether the EEG test is normal or abnormal. If it is abnormal, then the abnormalities are listed in order of importance and thus summarize the description section. Hence, mentions of the EEG activities and their attributes may be repeated, but with different words. The final section of the EEG report, provides the clinical correlations and explains what the EEG findings mean in terms of clinical interpretation (e.g. "very worrisome prognostic features").

Methods

The automatic annotation of the big data of EEG reports was performed by a Multi-task Active Deep Learning (MTADL) paradigm aiming to perform concurrently multiple annotation tasks, corresponding to the identification of (1) EEG activities and their attributes, (2) EEG events, (3) medical problems, (4) medical treatments and (5) medical tests mentioned in the narratives of the reports, along with their inferred forms of *modality* and *polarity*. When we considered the recognition of the modality, we took advantage of the definitions used in the 2012 i2b2 challenge[7] on

evaluating temporal relations in medical text. In that challenge, modality was used to capture whether a medical event discerned from a medical record actually happens, is merely proposed, mentioned as conditional, or described as possible. We extended this definition such that the possible modality values of "factual", "possible", and "proposed" indicate that medical concepts mentioned in the EEGs are actual findings, possible findings and findings that may be true at some point in the future, respectively. For identifying polarity of medical concepts in EEG reports, we relied on the same definition used in the 2012 i2b2 challenge, considering that each concept can have either a "positive" or a "negative" polarity, depending on any absent or present negation of its finding. Through the identification of modality and polarity of the medical concepts, we aimed to capture the neurologist's beliefs about the medical concepts mentioned in the EEG report. Some of the medical concepts mentioned in the EEG reports that describe the clinical picture of a patient are similar to those evaluated in the 2010 i2b2 challenge, as they represent medical problems, tests and treatments, thus we could take advantage of our participation in that challenge and use many of the features we have developed for automatically recognizing such medical concepts. However, EEG reports also contain a substantial number of mentions of EEG activities and EEG events, as they discuss the EEG test. The ability to automatically annotate all medical concepts from the EEGs entailed the development of an annotation schema that was created after consulting numerous neurology textbooks and inspecting a large number of EEG reports from the corpus. In fact, the development of the annotation schema represents the first step in our Multi-task Active Deep Learning (MTADL) paradigm, which required the following 5 steps:

STEP 1: *The development of an annotation schema;*
STEP 2: Annotation of initial training data;
STEP 3: Design of deep learning methods that are capable to be trained on the data;
STEP 4: Development of sampling methods for Multi-task Active Deep Learning system
STEP 5: Usage of the Active Learning system which involves:
    Step 5.a.: Accepting/Editing annotations of sampled examples
    Step 5.b.: Re-training the deep learning methods and evaluation the new system.

STEP 1: Annotation Schema: The annotation schema that we have developed considered EEG events, medical problems, treatment and tests to be annotated in similar ways as in the 2012 i2b2 challenge, namely by specifying (1) the boundary of each mention of concepts; (2) the concept type; (3) its modality and (4) its polarity. However, the EEG activities could not be annotated in the same way. First, we noticed that EEG activities are not mentioned in a continuous expression (see Example 1). To solve this problem, we annotated the *anchors* of EEG activities and their attributes. Since one of the attributes of EEG activities, namely, MORPHOLOGY, best defines these concepts, we decided to use it as an *anchor*. We considered three classes of attributes for EEG activities, namely (a) general attributes of the waves, e.g. the MORPHOLOGY, the FREQUENCY BAND; (b) temporal attributes and (c) spatial attributes. All attributes have multiple possible values associated with them. When annotating the MORPHOLOGY attribute we considered a hierarchy of values, distinguishing first two types: (1) Rhythm and (2) Transient. In addition, the Transient type contains three subtypes: Single Wave, Complex and Pattern. Each of these sub-types can take multiple possible values. An example of these annotations is provided in Example 2B.

STEP 2: Annotation of Initial Training Data: Initially, a sub-set of 39 EEG reports were manually annotated. The annotations were created by first running the medical concept recognition system reported in [7] to detect medical problems, tests, and treatments and their polarity and modality. The annotations that were obtained were manually inspected and edited, while also generating manual annotations for EEG Activities, their attributes as well as EEG

Attribute 1: **Morphology** ::= represents the type or "form" of EEG waves.
- Rhythm: continuous, rhythmic activity
- Transient
  - Single Wave:
    - V wave
    - Wicket spikes
    - Spike
    - Sharp wave
    - Slow wave
  - Complex: A sequence of two or more waves having a characteristic form or recurring with a fairly consistent form, distinguished from background activity.
    - K-complex
    - Sleep spindles
    - Spike-and-sharp-wave complex
    - Spike-and-slow-wave complex
    - Sharp-and-slow-wave complex
    - Triphasic wave
    - Polyspike complex
    - Polyspike-and-slow-wave complex
  - Pattern: any characteristic EEG Activity
    - Suppression
    - Amplitude Gradient
    - Slowing
    - Breach Rhythm
    - Benign Epileptic Transients of Sleep (BETS)
    - Photic driving (response)
    - Periodic Laterilized Epilepitiform Discharges (PLEDs)
    - Generalized periodic epileptiform discharges (GPEDs)
    - Epileptiform discharge (unspecified)
    - Disorganization
    - Positive occipital sharp transients of sleep (POSTS)
    - Unspecified: the default attribute value, used if no morphological information is given at all

Attribute 2:
Frequency Band
- Alpha (8 – 13 Hz)
- Beta (13 – 32 Hz)
- Delta (< 4 Hz)
- Theta (4 - 8 Hz)
- Gamma (> 32 Hz)

Attribute 3:
Background
- Yes
- No

Attribute 4: **Magnitude** ::= describes <u>the amplitude</u> of the EEG activity if it is emphasized in the EEG report • Low: e.g. subtle (spike), small (polyspike discharge) • High: e.g. high (voltage burst); high amplitude (spike); excess (theta) • Normal: the default value	Attribute 5: **Recurrence (<u>TEMPORAL</u>)** ::= describes <u>how often</u> the EEG activity occurs. • Continuous: the activity repeats in a continuous, uninterrupted manner • Repeated: the activity repeats intermittently • None: the activity occurs once
Attribute 6: **Dispersal (<u>SPATIAL</u>)** ::= describes the spread of the activity over regions of the brain • Localized (focal): limited to a small area of the brain • Generalized (diffuse): occurring over a large area of the brain or both sides of the head	Attribute 7: **Hemisphere (<u>SPATIAL</u>)** ::= describes which hemisphere of the brain the activity occur in. • Right • Left • Both

Location Attributes: **Brain Location (<u>SPATIAL</u>)** ::= describes the region of the brain in which the EEG activity occurs. The BRAIN LOCATION attribute of the EEG Activity indicates the location/area of the activity (corresponding to electrode placement under the standard 10-20 system).

- *Attribute 8:* Frontal (i.e. Anterior): Corresponds to the frontal region of the brain including all F*, Fp* and AF* electrodes
- *Attribute 9:* Occipital (i.e. Posterior): Corresponds to the occipital region of the brain including all O* electrodes
- *Attribute 10:* Temporal: Corresponds to the temporal region of the brain including all T* electrodes
- *Attribute 11:* Central: Corresponds to the central region of the brain including all C* electrodes
- *Attribute 12:* Parietal: Corresponds to the parietal region of the brain including all P* electrodes
- *Attribute 13:* Frontocentral: Corresponds to the area between the frontal and central regions of the brain including all FC* electrodes
- *Attribute 14:* Frontotemporal: Corresponds to the region between the frontal and temporal regions of the brain, including all FT* electrodes
- *Attribute 15:* Centroparietal: Corresponds to the region between the central and parietal regions of the brain including all CP* electrodes
- *Attribute 16:* Parieto-occiptal: Corresponds to the region between the parietal and occipital regions of the brain including all PO* electrodes

Events discovered in the sub-set of 39 EEG reports. The initial annotations represented the initial set of training data for two deep learning architectures, as illustrated in Figure 1.

STEP 3: Design of Deep Learning Architectures: The first architecture aims to identify (1) the *anchors* of all EEG activities mentioned in an EEG report; as well as (2) the boundaries of all mentions of EEG events, medical problems, medical treatments and medical tests. Examples of the annotation results of the first deep learning architecture are indexed with the two types discussed above in the following excerpt from an EEG report:

> Example 2A: *CLINICAL HISTORY: 58 year old woman found [unresponsive]$_2$, history of [multiple sclerosis]$_2$, evaluate for [anoxic encephalopathy]$_2$.*
> *MEDICATIONS: [Depakote]$_2$, [Pantoprazole]$_2$, [LOVENOX]$_1$.*
> *INTRODUCTION: [Digital video EEG]$_2$ was performed at bedside using standard 10.20 system of electrode placement with 1 channel of [EKG]$_2$. When the patient relaxes and the [eye blinks]$_2$ stop, there are frontally predominant generalized [spike and wave discharges]$_1$ as well as [polyspike and wave discharges]$_1$ at 4 to 4.5 Hz.*

Example 2A is an excerpt from of a typical EEG Report with several mentions of medical problems, tests, and treatments whose boundaries are denoted with brackets with subscript 1. Example 2A contains one EEG Event (subscript 1) and two EEG Activities (subscript 2). Like medical problems, tests, and treatments, EEG Events are identified by contiguous spans of text. However, unlike the other medical concept mentions, EEG Activities are often documented by multiple discontinuous spans of text. The two activities from example 2A are both observed frontally, have a generalized dispersal, and occur at a frequency of 4-4.5 Hz. Therefore we identify each EEG Activity by it's *anchor*, which is the span of text indicating the morphology of the activity, denoted in brackets with subscript 2 in Example 2B.

> Example 2B: *CLINICAL HISTORY: 58 year old woman found [unresponsive]$_{<TYPE=MP, MOD=Factual, POL=Positive>}$, history of [multiple sclerosis]$_{<TYPE=MP, MOD=Factual, POL=Positive>}$, evaluate for [anoxic encephalopathy]$_{<TYPE=MP, MOD=Possible, POL=Positive>}$.*
> *MEDICATIONS: [Depakote]$_{<TYPE=TR, MOD=Factual, POL=Positive>}$, [Pantoprazole]$_{<TYPE=TR, MOD=Factual, POL=Positive>}$, [LOVENOX]$_{<TYPE=TR, MOD=Factual, POL=Positive>}$.*
> *INTRODUCTION: [Digital video EEG]$_{<TYPE=Test, MOD=Factual, POL=Positive>}$ was performed at bedside using standard 10.20 system of electrode placement with 1 channel of [EKG]$_{<TYPE=Test, MOD=Factual, POL=Positive>}$. When the patient relaxes and the [eye blinks]$_{<TYPE=EV, MOD=Factual, POL=Positive>}$ stop, there are frontally predominant generalized [spike and wave discharges]$_{<MORPHOLGY=Transient>Complex>Spike\ and\ slow\ wave\ complex, FREQUENCYBAND=Delta, BACKGROUND=No, MAGNITUDE=Normal, RECURRENCE=Repeated, DISPERSAL=Generalized, HEMISPHERE=N/A, LOCATION={Frontal}, MOD=Factual, POL=Positive>}$ as well as [polyspike and wave discharges]$_{<MORPHOLGY=Transient>Complex>Polyspike\ and\ slow\ wave\ complex, FREQUENCYBAND=Delta, BACKGROUND=No, MAGNITUDE=Normal, RECURRENCE=Repeated, DISPERSAL=Generalized, HEMISPHERE=N/A, LOCATION={Frontal}, MOD=Factual, POL=Positive>}$ at 4 to 4.5 Hz.*

The annotations from Example 2B are produced by the second deep learning architecture, illustrated in Figure 1,
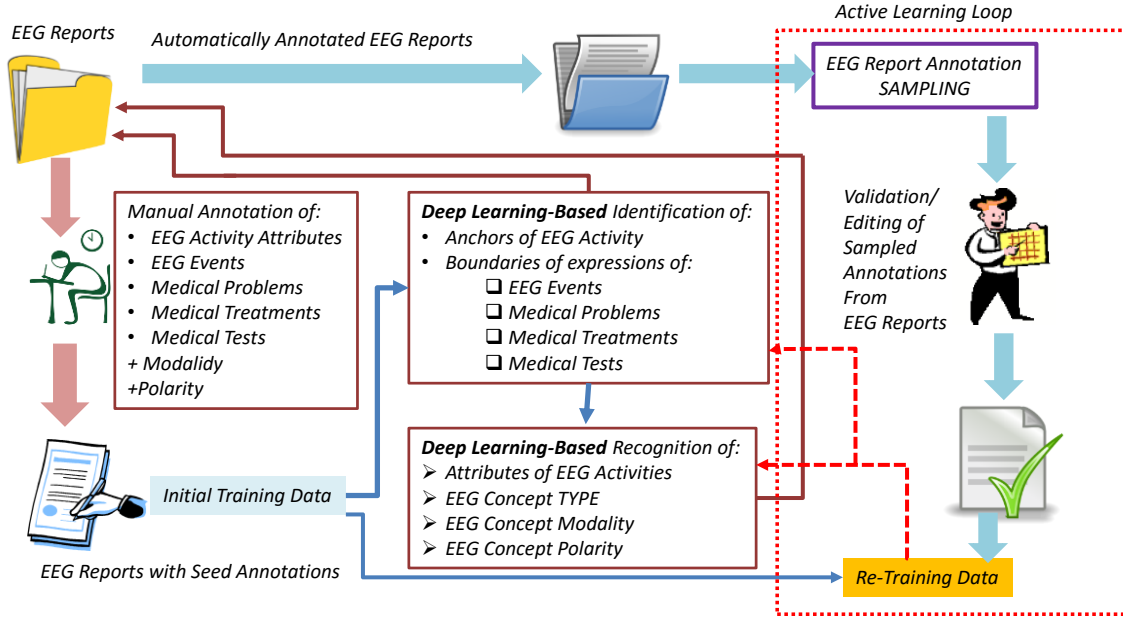
Figure 1: Architecture of the Multi-Task Active Deep Learning for annotating EEG Reports.

which is designed to recognize (i) the sixteen attributes that we have considered for each EEG activity, as well as (ii) the type of the EEG-specific medical concepts, discriminated as either an EEG event (EV), a medical problem (MP), a medical test (Test) or a medical treatment (TR). In addition, the second deep learning architecture identifies the modality and the polarity of these concepts.

After training the two deep learning architectures illustrated in Figure 1 on the initial training data obtained with manual annotations, we were able to automatically annotate the entire corpus of EEG reports. Because these automatically created annotations are not always correct, we developed an active learning framework to validate and edit these annotations, and provide new training data for the deep learning architectures.

STEP 4: Development of Sampling Methods: The choice of sampling mechanism is crucial for validation as it determines what makes one annotation a better candidate for validation over another. Multi-task Active Deep Learning (MTADL) is an active learning paradigm for multiple annotation tasks where new EEG reports are selected to be as informative as possible for *a set of annotation tasks* instead of a single annotation task. The sampling mechanism that we designed used the *rank combination* protocol[9], which combines several single-task active learning selection decisions into one. The usefulness score $s_{X_j}(\alpha)$ of each un-validated annotation α from an EEG Report is calculated with respect to each annotation task X_j and then translated into a rank $r_{X_j}(\alpha)$ where higher usefulness means lower rank (examples with identical scores get the same rank). Then, for each EEG Report, we sum the ranks of each annotation task to get the overall rank $r(\alpha) = \sum_{j=1} r_{X_j}(\alpha)$. All examples are sorted by this combined rank and annotations with lowest ranks are selected for validation. For each annotation task, we score an EEG Report d: $s_{X_j}(d) = \frac{1}{|d|}\sum_{a \in d} H(\alpha)$ where α is an annotation from d and $|d|$ is the number of annotations in document d, and $H(\alpha) = -\sum_c q_c^\alpha \log q_c^\alpha$ is the Shannon Entropy of α. This protocol favors selecting documents containing annotations the model is uncertain about from all annotation tasks.

STEP 5: Usage of the Multi-Task Active Deep Learning System: We performed several active learning sessions with our deep learning architectures. At each iteration, the deep learners are trained to predict annotations using the new validations. This process is repeated until (a) the error rate is acceptable; and (b) the number of validated examples is acceptable.

A. Feature Representations for Deep Learning operating on the EEG Big Data

The two deep learning architectures used in the Multi-task Active Deep Learning (MTADL) system illustrated in Figure 1 relied on two feature vector representations, that considered the features illustrated in Table 2. We used the GENIA tagger[10] for tokenization, lemmatization, Part of Speech (PoS) recognition, and phrase chunking. Stanford

Features used for **Deep Learning-Based** Identification of (a) Anchors of EEG Activity Attributes and (b) Boundaries of expressions of EEG Events, Medical Problems, Medical Treatments and Medical Tests	Features used for **Deep Learning-Based Recognition** of Attributes EEG Activities, EEG Concept TYPE, EEG Concept Modality and EEG Concept Polarity
1. The lemma of the token and the previous/next tokens 2. The PoS of the token and the previous/next tokens 3. The phrase chunk of the token and the previous/next tokens 4. The lemmas of the previous, current, and next tokens 5. The Brown cluster of the token 6. The UMLS Concept Unique Identifier (cui) of UMLS concepts containing the token 7. The title of the section containing the token	1. The medical concept mention itself 2. The lemmas of each token in the medical concept mention 3. The PoS of each token in the medical concept mention 4. The lemmas of 3 tokens before/after the medical concept mention 5. The title of the section containing the token *Context Features: For each token, **t**, in the sentence:* 6. The syntactic dependency path to **t**. 7. The number of words between the medical concept mention and **t** 8. The number of "hops" in the syntactic dependency path from the head of the medical concept mention to **t** 9. The number of medical concepts between the medical concept mention and **t**

CoreNLP was used for syntactic dependency parsing[11]. Brown Cluster[12] features generated from the entire TUH EEG corpus were used in both feature vector representations listed in Table 2. Brown clustering is an unsupervised learning method that discovers hierarchical clusters of words based on their contexts. We also used in the feature vector representation medical knowledge available from the Unified Medical Language System (UMLS)[13].

B. Deep Learning for Automatically Recognizing Medical concepts in EEG Reports

EEG reports mention multiple medical concepts in the narratives used in each report section. To find the spans of text that correspond to medical concepts, we trained two stacked Long Short-Term Memory (LSTM) networks[14]: one for detecting EEG Activity anchors and one for detecting the boundaries of all other medical concepts. For brevity, we will refer to both tasks as simply medical concept boundary detection in this subsection. The stacked LSTM networks process each document at the sentence level. To do this, we represent each sentence as a sequence of tokens $[w_1, w_2,..., w_N]$, and train both LSTMs to assign a label $b_i \in \{$ "I", "O", "B"$\}$ to each token w_i such that it will receive a label $b_i="B"$ if the token w_i is at the beginning of a mention of a medical concept, a label $b_i="I"$ if the token w_i is inside any mention of a medical concept and a label $b_i="O"$ if the token w_i is outside any mention of a medical concept.

For example, the token sequence "*occasional left anterior temporal [sharp and slow wave complexes]_{ACT}*" would correspond to the label sequence $[O,O, O,O,B,I,I,I]$, where tokens {*occasional, left, anterior, temporal*} are all assigned labels of *O*, as they are not part of the anchor of an EEG activity, although they describe its attributes, token {*sharp*} is assigned a label of *B*, and the tokens {*and, slow, wave, complexes*} are all assigned labels of *I*. This *IOB* notation allows medical concept mentions to be identified by continuous sequences of tokens starting with a token labeled *B* optionally followed tokens labeled *I*.
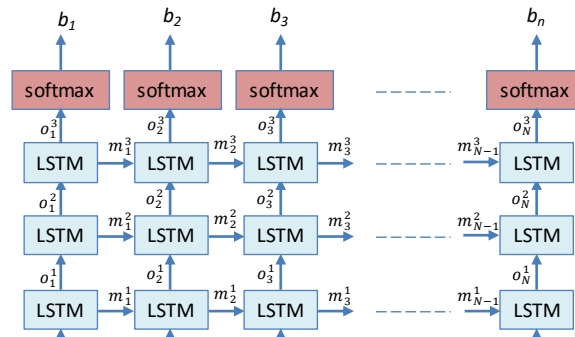
Figure 2: Deep Learning architecture for the identification of (1) the EEG activity anchors and (2) the boundaries of expressions of (a) EEG events, (b) medical problems; (c) medical tests and (d) medical treatments (e.g. medications).

To be able to use a deep learning architecture for automatically identifying the anchors of EEG activities and the boundaries of all other medical concepts in an EEG report, we first tokenized all reports, and represented each token w_i as a feature vector, t_i obtained by considering the features illustrated in Table 2. As illustrated in Figure 2, the features vectors $t_1, t_2, ..., t_N$ are provided as input to the stacked LSTMs to predict a sequence of output labels, $b_1, b_2, ..., b_N$. To predict each label b_i, the deep learning architecture considers (1) the vector representation of each token, t_i; as well as (2) the vector representation of all previous tokens from the sentence by updating a *memory* state that is shared throughout the network. LSTM cells also have the property that they can be "stacked" such that the outputs of cells on level l are used as the inputs to the cells on level on level $l + 1$. We used a stacked LSTM with 3 levels where the input to the first level is a sequence of token vectors and the output from the top level is used to determine the *IOB* labels for each token. The output from the top level, o_i^3, is a vector representing token w_i and every

previous token in the sentence. To determine the IOB label for token w_i, the output o_i^3 is passed through a softmax layer. The softmax layer produces a probability distribution over all IOB labels. This is accomplished by computing a vector of probabilities, q_i such that $q_{i,1}$ is the probability of label "I", $q_{i,2}$ is the probability of label "O", and $q_{i,3}$ is the probability of label "B". The predicted IOB label is then chosen as the label with highest probability, $y_i = \underset{j}{\text{argmax }} q_{ij}$. We use the same architecture to preform boundary detection for EEG Activity Anchors and all other medical concepts, but we train separate models for the two tasks.

C. Deep Learning with a ReLU Network for the Annotation of Attributes of EEG Activities, the Type of other Medical concepts and the recognition of Modality and Polarity in EEG Reports

In our annotation schema, we considered that each medical concept, a, is associated with a number of important attributes (16 attributes for EEG Activities as well as polarity and modality, and type, modality, and polarity for EEG events, medical problems, treatments and tests). After mentions of medical concepts have been automatically identified, we need to automatically determine each concept's attributes as well. Traditionally, attribute classification is performed by training a classifier, such as an SVM, to determine the value for each attribute. This approach would require training 18 separate attribute classifiers for EEG Activities and 3 classifiers for all other medical concepts. However, by leveraging the power of deep learning, we can simplify this task by creating one multi-purpose, high-dimensional vector representation of a medical concept, or *embedding*, and use this representation to determine each attribute simultaneously with the same deep learning network. Using a shared embedding allows important information to be shared between individual tasks. To accomplish this, we use the Deep Rectified Linear Network (DRLN) for multi-task attribute detection, illustrated in Figure 3.

Given a feature vector x_a representing a medical concept from an EEG report, based on the features from Table 2, the DRLN learns a *multi-task embedding* of the concept, denoted as e_a. To learn the multi-task embedding, the feature vector x_a is passed through 5 fully connected Rectified Linear Unit[15] (ReLU) layers. For $i \in 1, \cdots, 5$:
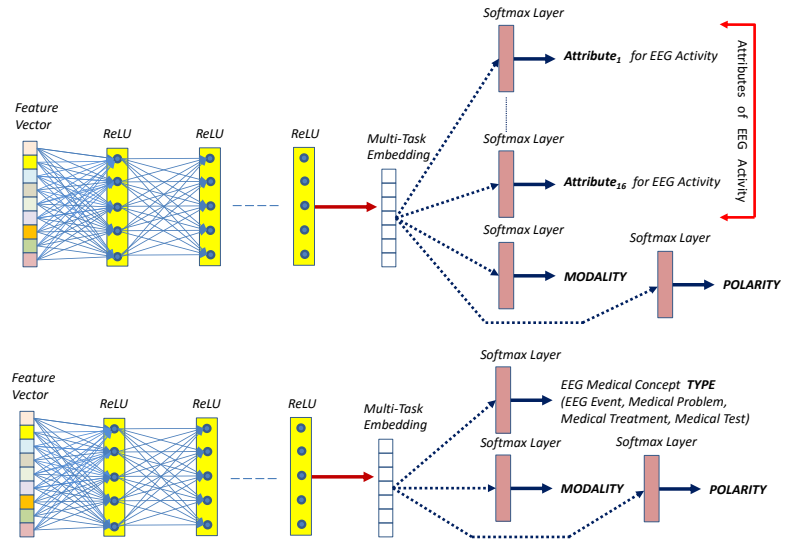
$$(9) \quad r_a^i = \max\{0, W^i \cdot r_a^{i-1} + b^i\}$$

where each W^i for $i \in 1, \dots, 5$ is a weight matrix, each b^i for $i \in 1, \dots, 5$ is a bias vector, r_a^1 is the input vector x_a, and r_a^5 is used as the multi-task embedding, e_a. The ReLU layers provide two major benefits that allow the network to function properly at depth: (1) ReLUs allow for a deep network configuration and (2) they

Figure 3: Deep Learning Architectures for Automatic Recognition of (1) attributes of EEG activities; (2) type for all the other medical concepts expressed in EEG reports; and (3) modality and polarity for all concepts.

learn sparse representations, allowing them to perform de facto internal feature selection[16]. The vanishing gradient problem effects deep networks by causing them to lose information used to update the weights in the network rapidly as the network gains depth[17], but ReLUs in particular avoid this problem.

As illustrated in Figure 3, the 16 attributes of the EEG activities are identified and annotated in the EEG reports by feeding the shared embedding into a separate softmax layer for each attribute. Formally, each softmax layer learns the predicted value $\tilde{y}_a^j$ for attribute j of medical concept a. Let q_a^j be the vector of probabilities produced by the softmax layer for attribute j of medical concept a. Each element q_{ak}^j of q_a^j is defined as:

$$(10) \quad \rho_a^j = \sigma(W_\rho \cdot e_a + b_\rho) \qquad\qquad (11) \quad q_{ak}^j = {e^{\rho_{ak}^j}}\Big/{\sum_{k'} e^{\rho_{ak'}^j}}$$

where the predicted attribute value $\tilde{y}_a^j = \underset{k}{\text{argmax}}\, q_{ak}^j$. Just as with boundary detection, we train two different networks, one for annotating the attributes, modality and polarity of EEG Activities and one for annotating the

Attributes & Attribute Values	A	P	R	F$_1$	#
Morphology	**0.990**	**0.757**	**0.704**	**0.724**	**1184**
DISORGANIZATION	0.979	0.887	0.788	0.834	80
GPEDS	0.999	0.000	0.000	0.000	1
POLYSPIKE_AND_WAVE	0.992	0.222	0.400	0.286	5
AMPLITUDE_GRADIENT	0.999	0.833	1.000	0.909	5
SPIKE_AND_SLOW_WAVE	0.995	0.941	0.970	0.955	66
SPIKE	0.992	0.850	0.708	0.773	24
PLEDS	0.993	0.750	0.500	0.600	12
LAMBDA_WAVE	1.000	1.000	1.000	1.000	18
K_COMPLEX	0.998	1.000	0.750	0.857	8
POLYSPIKE	0.991	0.750	0.529	0.621	17
SLOW_WAVE	0.994	0.941	0.923	0.932	52
RHYTHM	0.924	0.813	0.919	0.862	307
BETS	0.999	0.000	0.000	0.000	1
SLEEP_SPINDLE	0.998	1.000	0.913	0.955	23
SHARP_AND_SLOW_WAVE	0.996	0.600	0.500	0.545	6
SUPPRESSION	0.995	0.917	0.846	0.880	26
PHOTIC_DRIVING	0.998	1.000	0.947	0.973	38
TRIPHASIC_WAVE	0.999	1.000	0.909	0.952	11
SHARP_WAVE	0.989	0.886	0.963	0.923	81
WICKET	1.000	1.000	1.000	1.000	10
UNSPECIFIED	0.952	0.517	0.508	0.513	59
SPIKE_AND_SHARP_WAVE	1.000	0.000	0.000	0.000	0
EPILEPTIFORM_DISCHARGE	0.981	0.891	0.882	0.887	102
SLOWING	0.990	0.966	0.953	0.959	149
BREACH_RHYTHM	0.996	1.000	0.583	0.737	12
VERTEX_WAVE	1.000	1.000	1.000	1.000	30
Hemisphere	**0.924**	**0.775**	**0.754**	**0.762**	**1184**
*N/A	0.888	0.898	0.938	0.918	791
LEFT	0.942	0.717	0.711	0.714	121
RIGHT	0.965	0.756	0.782	0.768	87
BOTH	0.901	0.730	0.584	0.649	185
Magnitude	**0.909**	**0.806**	**0.710**	**0.750**	**1184**
HIGH	0.921	0.714	0.563	0.630	142
LOW	0.937	0.817	0.618	0.704	144
*NORMAL	0.869	0.886	0.950	0.917	898
Recurrence	**0.831**	**0.739**	**0.724**	**0.731**	**1184**
REPEATED	0.805	0.752	0.760	0.756	470
*NONE	0.787	0.750	0.773	0.761	520
CONTINUOUS	0.899	0.717	0.639	0.676	194
Dispersal	**0.871**	**0.775**	**0.733**	**0.751**	**1184**
LOCALIZED	0.882	0.759	0.684	0.720	263
*N/A	0.822	0.835	0.894	0.863	745
GENERALIZED	0.910	0.732	0.619	0.671	176
Frequency Band	**0.982**	**0.664**	**0.620**	**0.640**	**1184**
GAMMA	1.000	0.000	0.000	0.000	0
*N/A	0.945	0.940	0.983	0.961	811
DELTA	0.979	0.945	0.811	0.873	106
MU	1.000	0.000	0.000	0.000	0
ALPHA	0.981	0.897	0.870	0.883	100
BETA	0.992	0.957	0.918	0.937	73
THETA	0.975	0.910	0.755	0.826	94
BACKGROUND	**0.960**	**0.890**	**0.820**	**0.854**	**167**
LOCATION	**0.970**	**0.653**	**0.560**	**0.602**	**533**
PARIETO_OCCIPITAL	-	-	-	-	0
FRONTAL	0.929	0.724	0.640	0.679	139
OCCIPITAL	0.959	0.916	0.841	0.877	208
TEMPORAL	0.944	0.702	0.590	0.641	100
FRONTOTEMPORAL	0.993	0.727	0.615	0.667	13
FRONTOCENTRAL	0.990	0.882	0.789	0.833	38
CENTRAL	0.980	0.619	0.448	0.520	29
PARIETAL	0.995	0.000	0.000	0.000	6
CENTROPARIETAL	-	-	-	-	0
Polarity	**0.970**	**0.909**	**0.741**	**0.816**	**108**
Modality	**0.977**	**0.527**	**0.397**	**0.426**	**1178**
POSSIBLE	0.968	0.615	0.195	0.296	41
*FACTUAL	0.963	0.967	0.996	0.981	1136
PROPOSED	0.999	0.000	0.000	0.000	1

Table 2. Performance of our model when automatically detecting attributes of EEG activities. Default attribute values are denoted by an asterisk where applicable.

EEG Activity Anchors			Other Medical concept Boundaries		
Measure	**Exact**	**Partial**	**Measure**	**Exact**	**Partial**
P	.8949	.9591	**P**	.9161	.9469
R	.8125	.8228	**R**	.8797	.8831
F$_1$	.8517	.8857	**F$_1$**	.8975	.9139

Table 1. Performance of our model when automatically detecting anchors and boundaries of medical concepts

types, polarity and modality of all other medical concepts. EEG Activities have 18 attributes (the 16 EEG Activity specific attributes plust modality and polarity), therefore, the DRLN for learning EEG Activity attributes contains 18 softmax layers producing 18 predictions. In contrast, the DRLLN for learning the attributes of the other medical concepts has three attribute softmax layers, corresponding to (1) the type of the concept (EEG Event, medical problem, test, or treatment), (2) the modality, and (3) the polarity.

Results

In this section, we present and discuss the impact of applying Multi-task Active Deep Learning (MTADL) to the problem of detecting medical concepts and their attributes in EEG Reports. Specifically, we evaluated the performance of MTADL in terms of (1) the ability of the stacked LSTMs to detect the anchors of EEG Activities and the boundaries of all other medical concepts and; (2) the ability of the Deep ReLU Networks (DRLNs) to determine attributes of each medical concept (i.e. the attributes for EEG Activities, the type of all other medical concepts, and the modality and polarity of every medical concept). To measure the performance of our model when automatically detecting anchors and boundaries of medical concepts, we followed the evaluation procedure reported in the 2012 Informatics for Integrating Biology at the Bedside (i2b2) shared task[7]. We measured the precision (P), recall (R), and F$_1$ measure of the anchors and boundaries automatically detected by our system using 5-fold cross validation. As in [7], we report the performance of our model in terms of *exact* and *partial* matches. A predicted boundary is considered an exact match if it exactly matches any manually annotated boundary and it is considered a partial match it overlaps with any manually annotated boundary. Table 1 illustrates these results.

Clearly, our model is able to reliably identify both EEG Activity Anchors and other medical concept boundaries. It should be noted that the performance of detecting EEG Activity anchors was 5% lower than the performance of detecting the other medical concept boundaries. The difference in performance is not surprising given the fragmented nature of EEG Activity descriptions in EEG reports. While Table 1 shows the ability of our model to accurately determine anchors and boundaries of medical concepts, we were also interested in evaluating the automatically extracted attributes for each EEG Activity as well as the other medical concepts. For each attribute, we report the Accuracy (A), Precision (P), Recall (R), and F$_1$ measure for (1) each value as well as (2) the macro-average of all values for that attribute. Table 2 presents the performance of our first DRLN for determining the attributes of EEG Activities, while Table 3 presents the performance of our second DRLN for determining the attributes of other medical concepts. In both tables, we have also indicated the number of annotated mentions of each attribute as well as each of its values, indicated with the symbol '#'. Tables 1-3 show the promise of our model for detecting the anchors, boundaries, and attributes of

Attributes & Values	A	P	R	F$_1$	#
Concept Type	**0.970**	**0.943**	**0.936**	**0.939**	**2335**
TEST	0.983	0.982	0.958	0.970	669
PROBLEM	0.953	0.901	0.960	0.929	747
TREATMENT	0.971	0.964	0.898	0.930	500
EEG_EVENT	0.974	0.926	0.928	0.927	419
Modality	**0.973**	**0.742**	**0.605**	**0.659**	**2318**
POSSIBLE	0.977	0.634	0.406	0.495	64
FACTUAL	0.963	0.971	0.990	0.980	2199
PROPOSED	0.980	0.622	0.418	0.500	55
Polarity	**0.978**	**0.829**	**0.719**	**0.770**	**121**

Table 3. Performance of our model when automatically detecting attributes of EEG events and medical problems, treatments, and tests.

Figure 4. Learning curves for all annotations, shown over the first 100 EEG Reports annotated and evaluated with F$_1$ measure.

medical concepts from EEG reports. We also evaluated the impact of Multi-task Active Deep Learning (MTADL) on the performance of our model. Specifically, we measured the change in performance after each additional round of annotations. Figure 4 presents these results. Clearly the impact of MTADL on the performance of our model across all tasks was significant allowing it to achieve high performance after as few as 100 additional EEG Reports have been annotated.

Discussion

In general, it is clear that the DRLN was able to accurately determine the attributes of EEG, obtaining an overall accuracy of 93.8%. However, it is also clear that the model struggles to predict certain attribute values, for example MODALITY=Possible, MORPHOLOGY=Polyspike_and_wave, and BRAIN_LOCATION=Central. The degraded performance for these values is unsurprising as they are some of the least frequently annotated attributes in our data set (with 41, 5, 29 instances respectively). The difficulty of learning from a small number of annotations in the machine learning and natural language processing communities[18]. However, we believe that the performance of our model when detecting rare attributes could be improved in future work by incorporating knowledge from neurological ontologies[19] as well as other sources of general medical knowledge. We found that the performance of our DRLN for determining attribute of other medical concepts was highly promising, with an overall accuracy of 97.4%. However, we observed the same correlation between the number of annotations for an attribute's value and the DRLN's ability to predict that value. In the TUH EEG corpus, we found that nearly all mentions of EEG Events and medical problems, test, or treatments had a factual modality (96%). This follows the distribution of modality values reported in the 2012 i2b2 shared task (95% factual). The lowest performance of the DRLN was observed when determining the polarity attribute. The main source of errors for determining polarity was due to frequent ungrammatical sentences in the EEG Reports, e.g. "*There are rare sharp transients noted in the record but without after going slow waves as would be expected in epileptiform sharp waves*". We believe these errors could be overcome in future work by relying on parsers trained on medical data. As the MTDAL is being used, it enables us to generate EEG-specific qualified medical knowledge. We believe this knowledge can be enhanced by incorporating information from the EEG signals, creating a multi-modal medical knowledge representation. Such a knowledge representation is needed for reasoning mechanisms operating on big medical data.

Conclusion

In this paper we described a novel active learning annotation framework that operates on a big corpus of EEG Reports by making use of two deep learning architectures. The annotations follow a schema of semantic attributes characterizing EEG activities. Attributes define the morphology and magnitude as well as temporal (recurrence) and spatial (dispersal, brain location) characteristics of an EEG activity. The complex annotation schema enabled a Multi-task Active Deep Learning (MTADL) paradigm described in the paper. This paradigm uses one deep learning architecture based on two stacked LSTM networks to discover the textual boundaries of (a) EEG activity anchors and (b) expressions of EEG events, medical problems, tests, and treatments. After the anchors or boundaries are discovered, a second Deep Rectified Linear Network (DRLN) performs a multi-task attribute detection which identifies (a) any of the 16 attributes of EEG activities; and (b) the medical concept type which distinguishes between EEG event, medical problems, treatments, and test as well as (i) their modality and (ii) their modality. A crucial step in the MTADL paradigm is provided by the sampling mechanism for active learning. In this paper, we showed how instance sampling provided a significant increase in accuracy of annotation after each round of active learning. As the MTADL is being used, it enables us to generate an EEG specific medical knowledge that can be used to (1) improve patient cohort retrieval and (2) perform causal probabilistic inference.
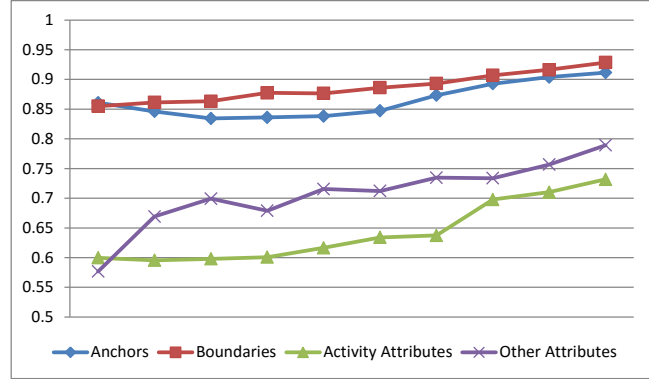
References

1. Smith SJM. EEG in the diagnosis, classification, and management of patients with epilepsy. J Neurol Neurosurg Psychiatry. 2005;76(suppl 2):ii2–7.
2. Tatum IV WO. Handbook of EEG interpretation [Internet]. Demos Medical Publishing; 2014. Available from: https://books.google.com/books?hl=en&lr=&id=BLsiAwAAQBAJ&oi=fnd&pg=PP1&dq=handbook+of+eeg +interpretation&ots=Zfq7Lw71LD&sig=9ZoI_ldUFBDHSKR1yOkZaWS4gIU
3. Beniczky S, Hirsch LJ, Kaplan PW, Pressler R, Bauer G, Aurlien H, et al. Unified EEG terminology and criteria for nonconvulsive status epilepticus. Epilepsia. 2013;54(s6):28–9.
4. Goodwin TR, Harabagiu SM. Multimodal Patient Cohort Identification from EEG Report and Signal Data. In: AMIA Annual Symposium Proceedings. American Medical Informatics Association; 2016.
5. Noachtar S, Binnie C, Ebersole J, Mauguiere F, Sakamoto A, Westmoreland B. A glossary of terms most commonly used by clinical electroencephalographers and proposal for the report form for the EEG findings. The International Federation of Clinical Neurophysiology. Electroencephalogr Clin Neurophysiol Suppl. 1999;52:21.
6. Hahn U, Beisswanger E, Buyko E, Faessler E. Active Learning-Based Corpus Annotation—The PathoJen Experience. In: AMIA Annual Symposium Proceedings [Internet]. American Medical Informatics Association; 2012 [cited 2016 Sep 23]. p. 301. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3540513/
3. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc JAMIA. 2013 Sep;20(5):806–13.
8. Kale DC, Che Z, Bahadori MT, Li W, Liu Y, Wetzel R. Causal Phenotype Discovery via Deep Networks. AMIA Annu Symp Proc. 2015 Nov 5;2015:677–86.
9. Reichart R, Tomanek K, Hahn U, Rappoport A. Multi-Task Active Learning for Linguistic Annotations. In: ACL [Internet]. 2008 [cited 2016 Sep 22]. p. 861–9. Available from: http://www.anthology.aclweb.org/P/P08/P08-1.pdf#page=905
10. Tsuruoka Y, Tateishi Y, Kim J-D, Ohta T, McNaught J, Ananiadou S, et al. Developing a robust part-of-speech tagger for biomedical text. In: Panhellenic Conference on Informatics [Internet]. Springer; 2005 [cited 2016 Sep 22]. p. 382–92. Available from: http://link.springer.com/chapter/10.1007/11573036_36
11. Chen D, Manning CD. A Fast and Accurate Dependency Parser using Neural Networks. In: EMNLP [Internet]. 2014 [cited 2016 Sep 22]. p. 740–50. Available from: http://www-cs.stanford.edu/~danqi/papers/emnlp2014.pdf
12. Brown PF, Desouza PV, Mercer RL, Pietra VJD, Lai JC. Class-based n-gram models of natural language. Comput Linguist. 1992;18(4):467–79.
13. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(suppl 1):D267–70.
14. Pascanu R, Gulcehre C, Cho K, Bengio Y. How to construct deep recurrent neural networks. ArXiv Prepr ArXiv13126026 [Internet]. 2013 [cited 2016 Sep 22]; Available from: http://arxiv.org/abs/1312.6026
15. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10) [Internet]. 2010 [cited 2016 Sep 22]. p. 807–14. Available from: http://machinelearning.wustl.edu/mlpapers/paper_files/icml2010_NairH10.pdf
16. Glorot X, Bordes A, Bengio Y. Deep Sparse Rectifier Neural Networks. In: Aistats [Internet]. 2011 [cited 2016 Sep 22]. p. 275. Available from: http://www.jmlr.org/proceedings/papers/v15/glorot11a/glorot11a.pdf
17. Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw. 1994;5(2):157–66.
18. Martin JH, Jurafsky D. Speech and language processing. Int Ed [Internet]. 2000 [cited 2016 Sep 23];710. Available from: http://www.ulb.tu-darmstadt.de/tocs/203636384.pdf
19. Sahoo SS, Lhatoo SD, Gupta DK, Cui L, Zhao M, Jayapandian C, et al. Epilepsy and seizure ontology: towards an epilepsy informatics infrastructure for clinical research and patient care. J Am Med Inform Assoc. 2014;21(1):82–9.