Clinical Natural Language Processing with Deep Learning

Sadid A. Hasan and Oladimeji Farri

Abstract The emergence and proliferation of electronic health record (EHR) systems has incrementally resulted in large volumes of clinical free text documents available across healthcare networks. The huge amount of data inspires research and development focused on novel clinical natural language processing (NLP) solutions to optimize clinical care and improve patient outcomes. In recent years, deep learning techniques have demonstrated superior performance over traditional machine learning (ML) techniques for various general-domain NLP tasks e.g. language modeling, parts-of-speech (POS) tagging, named entity recognition, paraphrase identification, sentiment analysis etc. Clinical documents pose unique challenges compared to general-domain text due to widespread use of acronyms and non-standard clinical jargons by healthcare providers, inconsistent document structure and organization, and requirement for rigorous de-identification and anonymization to ensure patient data privacy. This tutorial chapter will present an overview of how deep learning techniques can be applied to solve NLP tasks in general, followed by a literature survey of existing deep learning algorithms applied to clinical NLP problems. Finally, we include a description of various deep learning-driven clinical NLP applications developed at the Artificial Intelligence (AI) lab in Philips Research in recent years - such as diagnostic inferencing from unstructured clinical narratives, relevant biomedical article retrieval based on clinical case scenarios, clinical paraphrase generation, adverse drug event (ADE) detection from social media, and medical image caption generation.

Sadid A. Hasan
Artificial Intelligence Lab, Philips Research North America, Cambridge, MA, USA. e-mail: sadid.hasan@philips.com

Oladimeji Farri
Artificial Intelligence Lab, Philips Research North America, Cambridge, MA, USA. e-mail: dimeji.farri@philips.com

1 Introduction

Over the ages, humans continuously use written and spoken language as a means of expressing and communicating our conceptualization of abstract and real-life scenarios of varying complexity. Documented narratives are viewed as essential sources of knowledge that can be transferred and synthesized to retrieve pertinent insights for decision-making across all domains of expertise. The explosive growth and access to unstructured data in the digital universe since the birth of the internet has helped establish natural language processing (NLP) as one of the most important technologies needed to address complex and knowledge-dependent tasks such as automated search, machine translation, automated question answering, and opinion mining. In particular, the emergence of electronic health record (EHR) systems since the 1960s has incrementally resulted in large volumes of clinical free text documents available across healthcare networks, with the huge amount of data inspiring research and development focused on novel clinical NLP solutions to optimize clinical care and improve patient outcomes across the care continuum [1].

In recent years, deep learning techniques have demonstrated superior performance over traditional machine learning (ML) techniques for various general-domain NLP tasks e.g. language modeling, parts-of-speech (POS) tagging, named entity recognition, paraphrase identification, and sentiment analysis. Clinical documents generally pose unique challenges compared to general-domain text due to widespread use of acronyms and non-standard clinical jargons by healthcare providers, inconsistent document structure and organization, and requirement for rigorous de-identification and anonymization to ensure patient data privacy. Ultimately, overcoming these challenges could foster more research and innovation for various useful clinical applications including clinical decision support, patient cohort identification, patient engagement support, population health management, pharmacovigilance, personalized medicine, and clinical text summarization.

This tutorial chapter is an overview of how deep learning techniques can be applied to solve NLP tasks, followed by a literature survey of existing deep learning algorithms applied to clinical NLP problems, and finally, a detailed description of various deep learning-driven clinical NLP applications developed at the Artificial Intelligence lab in Philips Research in recent years - such as diagnostic inferencing from unstructured clinical narratives, relevant biomedical article retrieval based on clinical case scenarios, clinical paraphrase generation, adverse drug event (ADE) detection from social media, and medical image caption generation.

2 Deep Learning for NLP

NLP is a field intersecting computer science, artificial intelligence, and linguistics where the goal is to process and understand human language to perform useful tasks (e.g. automated question answering, language translation). NLP is generally considered to be an AI-complete problem due to various complexities involved in repre-

senting, learning, and using linguistic, situational, world or visual knowledge. Given an input text, NLP typically involves processing at various levels such as tokenization, morphological analysis, syntactic analysis, semantic analysis, and discourse processing.

Deep learning is a type of machine learning technique that utilizes multi-layered (hence the term *deep*) neural network architectures to learn hierarchical representations of data. Traditional machine learning approaches require labor-intensive feature engineering for data representation [2]. By contrast, deep learning approaches can automatically learn multiple levels of representations with increasing order of abstractions [3]. Figure 1 shows an example deep neural network architecture. The recent surge in deep learning can be credited to the following: the availability of a large amount of unlabeled data as well as faster computing resources with powerful graphics processing units (GPUs), development of new algorithms and frameworks, and easier adaptations/transformations of learned features/representations from data to a related or a new domain of interest (transfer learning).

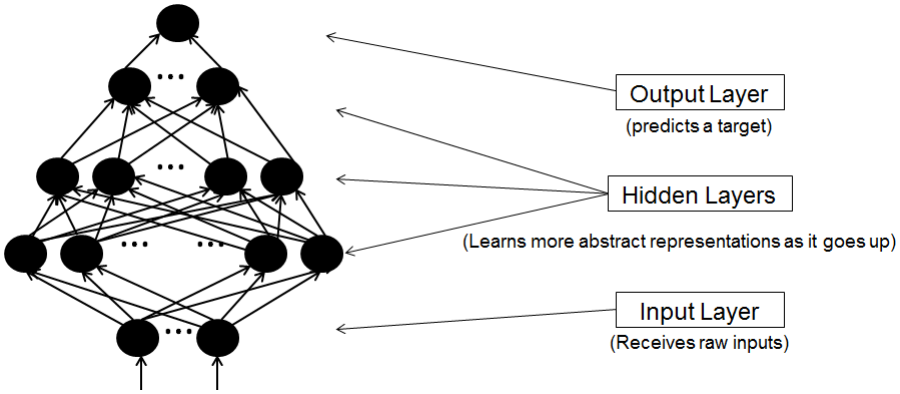

Fig. 1: A deep neural network architecture.

Deep learning typically works well to solve non-linear classification problems with naturally occurring hierarchical inputs such as language and images. In recent years, non-linear neural network models applied to NLP techniques have achieved promising results over approaches that use linear models such as support vector machines (SVMs) or logistic regression [4].

In this section, we will introduce how deep learning techniques can be applied to solve NLP problems in general. First, we will provide a brief description of how input representations are generated for NLP applications. Then, we will focus on two deep learning architectures that are widely used by the NLP research community: convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Finally, we will describe memory networks and deep reinforcement learning to facilitate the understanding of clinical NLP applications discussed in Section 3.2.

2.1 Input Representation

Natural language inputs are typically represented as features such as words, named entities, and parts-of-speech tags. Bag-of-Words (BOW) modeling or one-hot vector encoding techniques can be used to represent the meaning of the words in a given text. In BOW modeling, the presence or absence of a word in a sentence compared to the underlying corpus can be used to create a fixed length vector representation. Alternatively, term frequency-inverse document frequency (TF-IDF) scoring can be used to create vector representations of input text. In one-hot vector encoding, each word can be represented as a vector of size n, where n stands for the dimensionality of the vector denoting the number of words present in the corpus/vocabulary. For example, if there are 10 words in the vocabulary, each word can be represented as a 10-dimensional vector with one specific position set to 1 and the rest to 0. The main limitations of BOW and one-hot encoding approaches include inconsideration of word orders, dependency of dimensionality on the vocabulary size, and consequently, sparsity [4, 5].

Distributional similarity-based representations can be used to alleviate some of the aforementioned limitations by forming a window-based co-occurrence matrix[1] for an underlying corpus. However, there still remain dimension size and sparsity related issues, which can be alleviated further by reducing the dimensions via techniques such as singular value decomposition (SVD) [6]. But, SVD involves higher computational cost with difficulty to include new words/documents into the considered corpus. A solution to this is to directly learn low-dimensional word vectors from the corpus. Instead of computing co-occurrence counts, the main idea here is to either predict surrounding words in a certain window of each word (Skip-gram model) or predict each word given the surrounding words (Continuous BOW or CBOW model) to represent words in terms of vectors (Word2Vec) [7]. A feed-forward neural network architecture can be used to learn the vector representations from a corpus by minimizing a loss function such as hierarchical softmax, cross entropy, negative sampling etc. using an optimization technique such as stochastic gradient descent (SGD) [8].

Deep learning for NLP applications mainly rely on learning high-dimensional vector representations of character-level n-grams, words, phrases, sentences, or documents and their relationships (called *embeddings*) using deep neural network architectures [5, 9]. The trained language model transforms semantically similar textual units into similar vector representations [10, 8]. The main advantage of such architecture over the traditional bag-of-words model is its ability to capture the embedded ordering and semantics by learning fixed-length vector representations for variable-length text structures (via neural network architectures like RNNs), thereby allowing the training of generative models for complex NLP tasks such as machine translation and dialogue generation.

[1] This matrix can be constructed based on simple frequency count of co-occurring words in a fixed window size across all possible combinations of the words in a corpus. The matrix can be plotted in a multi-dimensional space to essentially group the words with similar co-occurrence values together denoting their semantic and syntactic relationships.

2.2 Convolutional Neural Networks (CNNs)

CNN is a multi-layer neural network that uses a special kind of linear mathematical operation called *convolution* instead of general matrix multiplication in at least one of its layers. CNNs automatically learn the values of the filters (a.k.a. kernels) from the input data based on an underlying task. Each filter essentially encodes a local view of lower-level features into a higher-level representation via operating a sliding window function to the input. Typically, a CNN is composed of 3 layers/stages: convolution, detection (nonlinear activation), and pooling - to portray two important aspects: location invariance (considers the presence of a feature as important, not the specific location) and local compositionality (encodes lower-level features into higher-level representations as they are passed to higher layers) [3]. The convolution layer applies several convolutions parallelly to generate corresponding linear activations and then, the detector layer applies a nonlinear activation function to each linear activation. The pooling layer computes the maximum value (max-pooling) or the average value (average-pooling) of a subset of outputs from the underlying layers in order to provide it as input to the higher layers. Stacks of convolutional and pooling layers can be added on top of a pooling layer to construct a deep convolutional neural network. Figure 2 shows a simple CNN architecture. $W_1, W_2, \cdots, W_6$ are the weights of the model, and shared weights are shown with the same color. Note that, convolution and detection are plotted together in the figure using rectified linear unit (ReLU) symbols in the convolution layer nodes.

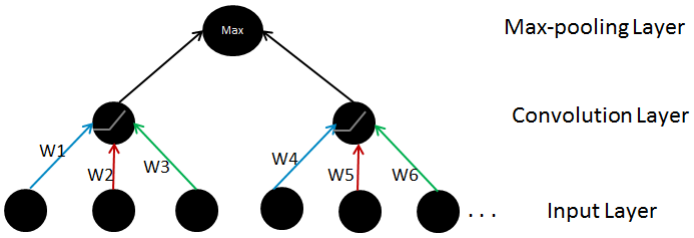

Fig. 2: A simple CNN architecture.

In the NLP domain, CNNs are generally shown to be effective in solving classification tasks [11] such as sentiment analysis, spam detection or topic categorization because they work similar to the BOW principles (i.e. *location invariance* being similar to the lack of consideration of word order). Multiple filters/kernels can be applied to learn various features from the input data. Each filter can essentially transform a set of words in a certain window of size k to a d-dimensional (each dimension is also known as a *channel*) vector representation that embeds key aspects of the words in consideration [12]. Different filters can focus on certain words inside variable window sizes to capture different features from the corpus. For example, in a sentiment analysis task, a filter can detect a negation feature e.g. “not amazing”

from the sentence "the product is not amazing". However, since CNNs do not capture the global information from the sentence due to location invariance and local compositionality properties, they are not able to distinguish the difference between "not amazing" and "amazing not (so much)". Hyperparameters of a CNN model include number of filters, convolution type (narrow vs. wide), stride size[2], and number of channels.

Let $x_i(t)$ be the input vector (which can be pre-trained on a large unlabeled dataset or can simply be initialized as one-hot encodings) for the i-th word $w_i(t)$ of input sentence s, W be the corresponding weight matrix called kernel/filter, b be the bias vector, and σ be the component-wise non-linear activation function; then a computational unit of the convolutional layer associated with the i-th word can be formulated as follows:

$$\sigma(W \cdot x_i(t) + b) \tag{1}$$

W, and b are the parameters of the model that are learned through training on a labeled data set and can be shared across all neurons of the same layer. The *rectifier*, $\sigma(x) = max(0,x)$ can be used as the non-linear activation function (other non-linear activation functions include hyperbolic tangent or tanh(x)), max-pooling for computing higher-layer abstractions, and stochastic gradient descent for optimization where the objective is to minimize the square loss or cross-entropy loss with respect to the labeled training set. Finally, the output layer of the network may use a linear classifier that exploits the learned features to predict the label for any classification task [11, 13, 14].

In contrast to RNNs (discussed in the next subsection) that maintain a hidden state to encode the previous sequence of the input data, CNNs do not rely on the past steps to allow parallel processing of input elements for faster computations. Thereby, CNNs are recently shown to achieve state-of-the-art results in sequence to sequence learning tasks e.g. neural machine translation, at a faster speed compared to RNN-based models [15, 16].

2.3 Recurrent Neural Networks (RNNs)

RNNs generally work well for modeling sequences. Hence, they are used to solve various NLP tasks due to their ability to deal with variable-length input and output [17]. The RNN network architecture is similar to the standard feedforward neural network with the exception that hidden unit activation at a particular time t is dependent on that of time $t-1$.

Figure 3 shows an unrolled RNN architecture, where x_t, y_t, h_t are the input, output, and hidden state at time step t, and W, U, V are the parameters of the model corresponding to *input*, *hidden*, and *output* layer weights (shared across all time steps).

[2] Stride size denotes the amount by which a filter is shifted across the input data.

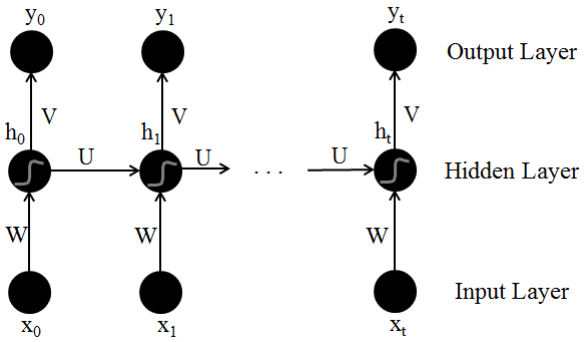

Fig. 3: Generic recurrent neural network architecture.

The hidden state h_t is essentially the memory of the network as it can capture necessary information about an input sequence by exploiting the previous hidden state h_{t-1} and the current input x_t as follows:

$$h_t = f(Wx_t + Uh_{t-1}), \tag{2}$$

where f is an element-wise nonlinear activation function. The output y_t is computed similarly as a function of the memory at time t: Vh_t. Although RNN is theoretically a powerful model to encode sequential information, in practice it often suffers from the vanishing/exploding gradient problems while learning long-range dependencies [18]. Long Short-Term Memory (LSTM) networks [19] and Gated Recurrent Units (GRU) [20] are known to be successful remedies to these problems.

A LSTM unit basically computes the hidden state h_t using a different approach than the generic RNN framework by introducing a gating mechanism. The main idea is to control how much information to keep from the old memory and the most recent information. Formally, LSTM computes h_t using the following equations:

$$\begin{aligned}
i_t &= \sigma(W^i x_t + U^i h_{t-1}) \\
f_t &= \sigma(W^f x_t + U^f h_{t-1}) \\
o_t &= \sigma(W^o x_t + U^o h_{t-1}) \\
g_t &= \tanh(W^g x_t + U^g h_{t-1}) \\
c_t &= c_{t-1} \odot f_t + g_t \odot i_t \\
h_t &= \tanh(c_t) \odot o_t
\end{aligned} \tag{3}$$

where i_t, f_t, o_t are the *input*, *forget* and *output* gates, g_t is the candidate hidden state, $\sigma(.)$ and $\tanh(.)$ are the element-wise sigmoid and hyperbolic tangent functions, and $\odot$ denotes element-wise multiplication. Note that, all three gates and the

candidate hidden state are computed in a similar fashion as Eq. 2 with different weight parameters. c_t is the internal memory state that is essentially computed based on the previous memory state at time $t-1$ and the new input information at time t. Finally, h_t is calculated by combining the memory with the output gate, which determines how much of the internal state information needs to be passed along to the higher layers of the network.

GRU is a simplified version of LSTM with less number of parameters per unit, and thus, the total number of parameters can be greatly reduced for a large neural network [20]. In contrast to LSTM, GRU does not have an internal memory state and the output gate; rather it introduces two gates termed *update* and *reset* to accomplish the same goal. In fact, GRU computes the hidden state h_t in a slightly alternative fashion as follows:

$$\begin{aligned} z_t &= \sigma(W^z x_t + U^z h_{t-1}) \\ r_t &= \sigma(W^r x_t + U^r h_{t-1}) \\ k_t &= \tanh(W^k x_t + U^k (r_t \odot h_{t-1})) \\ h_t &= (1 - z_t) \odot k_t + z_t \odot h_{t-1} \end{aligned} \tag{4}$$

where z_t, r_t are the update gate and the reset gate, and k_t is the candidate hidden state. Note that, z_t, r_t are computed similarly as LSTM (using different weight parameters) where z_t determines how much of the old memory to keep while r_t denotes how much new information is needed to be combined with the old memory. Finally, k_t is computed by exploiting r_t and h_t is calculated to denote the amount of information needed to be transmitted to the following layers.

2.4 Memory Networks (MemNNs)

MemNNs are a class of models that contain an external memory and a controller to read from and write to the memory [21, 22]. MemNNs read a given input source and a knowledge source several times (hops) while updating an internal memory state. The memory state is the representation of relevant information from a knowledge source optimized to solve a given task. In particular, a MemNN stores all information (e.g. knowledge base, background context) into the external memory, assigns a relevance probability to each memory slot using content-based addressing schemes, and reads contents from each memory slot by taking their weighted sum. MemNNs are generally harder to train than traditional networks as they need supervision at every layer and they do not scale easily to a large memory. End-to-End Memory Networks [21] and Key-Value Memory Networks (KV-MemNNs) [23] can alleviate these issues by training multiple hops over memory (allowing for less supervision) and compartmentalizing memory slots into hashes.

The basic structure of a MemNN involves learning memory representations from a given knowledge base. Memory is typically organized as t number of slots, $m_1, \ldots m_t$. For a given input text $x_1, \ldots, x_n$, an external knowledge base represented as key-value pairs $(k_1, v_1), (k_2, v_2), \ldots, (k_m, v_m)$, and the ground truth outputs y, a model $\mathscr{F}$ can be learned as the following:

$$\mathscr{F}(x_n, (k_m, v_m)) = \hat{y} \rightarrow y \tag{5}$$

where the function $\mathscr{F}$ has the following parts *I (Input memory representation), G (Generalization), O (Output memory representation), and R (Response)* which are the standard components of MemNNs [22].

2.5 Deep Reinforcement Learning

Reinforcement learning is a machine learning technique that considers an agent to learn to take actions in an environment such that it can achieve the maximum possible reward in the future. The environment can be modeled as a Markov Decision Process (MDP) that includes a set of states, a set of actions, a transition function to model the probability to move from one state to the other after taking an action, and a reward function that assigns a reward to the agent after its transition to a new state. In a state s, the agent takes an action a to get to the next state, $s' = s + a$. A reward function $r(s,a)$ can be used to estimate the reward at each state s after taking an action a. A reinforcement learning problem can be formulated by estimating a state-action value function $Q(s,a)$, which determines the optimal action a to take in a state s using the Q-learning technique [24]. In order to learn the Q-value, the iterative updates are derived from the Bellman equation [25]:

$$Q_{i+1}(s,a) = E[r + \gamma max_{a'} Q_i(s', a') | s, a], \tag{6}$$

where γ is a discount factor for the future rewards and the expectation is over the whole training process. It is impractical to maintain the Q-values for all possible state-action pairs. Hence, the Q-function can be approximated using a deep Q-network (DQN) architecture [26] that uses a deep neural network (hence called *deep reinforcement learning*) to obviate the need of explicitly designing the state and action space. The DQN architecture approximates the Q-value function and predicts $Q(s,a)$ for all possible actions.

3 Clinical NLP with Deep Learning

In this section, we will focus on the application of deep learning techniques for clinical NLP problems. First, in Section 3.1 we will discuss the most notable recent clinical NLP applications developed by the research community that leverage deep

learning. Then, in Section 3.2 we will describe some deep learning-driven clinical NLP applications developed at the AI lab in Philips Research.

3.1 Literature Survey

CNNs have been successfully applied to a variety of biomedical NLP tasks in the literature. For example, CNNs are effectively used to build a biomedical article classification model to identify the hallmarks of cancer associated with a given article abstract [27], to learn time expression representation for clinical temporal relation extraction [28], to model the article relevance with respect to the query for the task of biomedical article retrieval [29], to identify protein-protein interaction relations from biomedical articles [30], to extract drug-drug interactions with an attention mechanism [31], to classify radiology free text reports based on pulmonary embolism findings [32], to classify patient portal messages towards providing necessary support [33], and to recognize named entities from biomedical text [34]. CNN-based models are also shown to achieve better performance over the traditional machine learning classifiers for automated coding of radiology reports using the International Classification of Diseases (ICD-10) coding scheme [35]. Inspired by the aforementioned success of CNNs for various clinical NLP applications, we proposed a novel semi-supervised CNN architecture (discussed in Section 3.2.4) for automated ADE detection in social media. Unlike conventional systems [36, 37, 38, 39, 40, 41] that typically use lexicon- and traditional machine learning-based approaches relying on expert annotations to generate large amounts of labeled data to train supervised machine learning models for ADE detection, our proposed system can efficiently learn from large volumes of unlabeled data in combination with a relatively small *seed set* of labeled ADEs.

Some recent works explore the use of RNN architectures for the task of detecting clinical events such as disorders, treatments, tests, and adverse drug events from free text EHR notes [42, 43, 44], and for de-identification of patient data in EHRs [45, 46, 47]. Bidirectional RNNs/LSTMs are used to develop models for missing punctuation prediction in medical reports [48], for the task of biomedical events trigger identification [49], to model relational and contextual similarities between the named entities in biomedical articles to understand meaningful insights towards providing appropriate treatment suggestions [50], to extract clinical concepts from EHR reports [51], and for named entity recognition from clinical text [52, 53]. A recent work builds a bidirectional LSTM transducer by leveraging knowledge graph embeddings to detect adverse drug reaction in social media data [54]. RNNs are also used in combination with CNNs to learn disease name recognition models with word- and character-level embedding features [55]. Motivated by these prior works, we proposed an attention-based bidirectional RNN architecture inside an encoder-decoder framework for the task of clinical paraphrase generation (discussed in Section 3.2.3) by casting it as a monolingual neural machine translation problem. Unlike earlier work on clinical-domain specific paraphrasing that

uses some unsupervised textual similarity measures to generate/extract lexical and phrasal paraphrases from monolingual parallel and comparable corpora [56, 57], or adopts a semi-supervised word embedding model for medical synonym extraction [58], our work was the first to propose a neural network-based architecture that can model word/character sequences to essentially address all granularities of paraphrase generation [59] for the clinical domain [60]. Furthermore, we have leveraged the strengths of deep CNNs and attention-based RNNs in an encoder-decoder framework to train medical image caption generation models (discussed in Section 3.2.5) that achieved superior results in a benchmark evaluation challenge.

As stated in Section 2.4, variants of memory networks provide flexibility to leverage knowledge sources to effectively accomplish NLP tasks requiring complex reasoning and inferencing e.g. question answering. In this regard, we proposed a novel condensed memory network architecture for the task of clinical diagnostic inferencing from unstructured clinical text narratives (see Section 3.2.1 for details). Unlike conventional clinical decision support (CDS) systems that leverage LSTM neural networks trained on time series data for diagnosis classification [61, 62], our work was the first to propose the use of a novel memory network model trained on unstructured clinical texts to recommend differential diagnoses. We have also utilized key-value memory networks for clinical diagnostic inferencing as a core component of our biomedical article retrieval system discussed in Section 3.2.2.

Existing applications of reinforcement learning for related CDS tasks mainly focus on modalities like medical imaging [63] or specific domain-dependent use cases, and clinical trials [64, 65, 66]. Some prior works demonstrate the utility of deep reinforcement learning techniques for challenging tasks like playing games and entity extraction [26, 67, 68]. These works inspired us to propose novel deep reinforcement learning-based algorithms for clinical diagnosis inference from unstructured text narratives (discussed in Section 3.2.1).

3.2 Applications Developed in Philips Research

3.2.1 Diagnostic Inferencing

Clinicians perform complex cognitive processes to infer the probable diagnosis after observing several variables such as the patient's past medical history, current condition, and various clinical measurements. The cognitive burden of dealing with complex patient situations could be reduced by having an automated assistant provide suggestions to physicians of the most probable diagnoses for optimal clinical decision-making.

We explored the discriminatory capability of the unstructured free-text clinical notes to correctly infer the most probable diagnoses from a complex clinical scenario [69]. We also explored the use of an external knowledge source like Wikipedia from which the model can extract relevant information, such as signs and symptoms for various diseases. Our main goal was to combine such an external clinical

knowledge source with the free-text clinical notes and use the learning capability of memory networks to correctly infer the most probable diagnoses.

For real world tasks, a large amount of memory is required to achieve state-of-the-art results. Following the effective use of memory networks in solving question answering tasks, we introduced Condensed Memory Networks (C-MemNNs), an approach to efficiently store condensed representations in memory, thereby maximizing the utility of limited memory slots. We showed that a condensed form of memory state which contains some information from earlier hops learns efficient representation. We took inspiration from human memory retention patterns for this model. Humans can learn new information and yet retain relatively older memories as abstractions. We formulated the clinical diagnostic inferencing problem as a supervised multi-label multi-class classification problem using C-MemNNs. Figure 4 demonstrates the iterative updating process of the condensed memory state (left) and the overall condensed memory network architecture (right) for clinical diagnostic inferencing. Interested readers are referred to [69] for in-depth details.

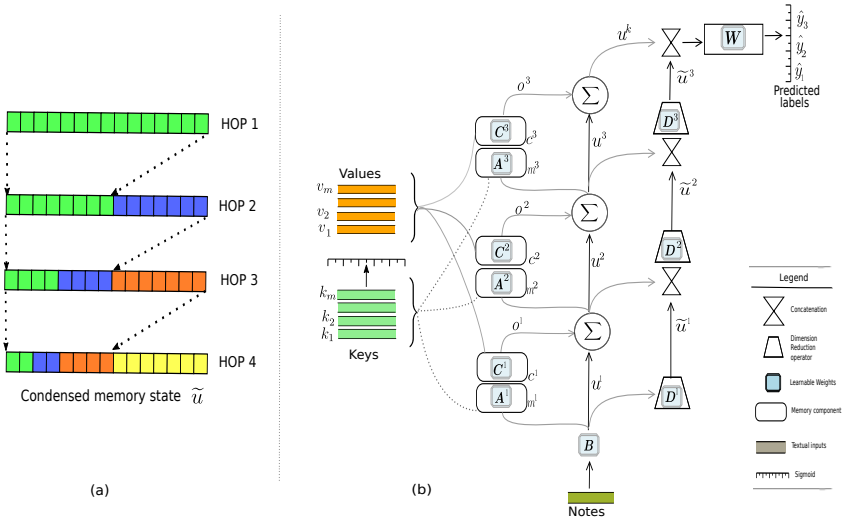

Fig. 4: Condensed memory networks for clinical diagnostic inferencing [69].

MIMIC−III ((Medical Information Mart for Intensive Care III)) [70], a large freely-available clinical database, was used for our experiments. It contains physiological signals and various measurements captured from patient monitors, and comprehensive clinical data obtained from hospital medical information systems for over 58*K* Intensive Care Unit (ICU) patients. We used the *noteevents* table from MIMIC−III: v1.3, which contains unstructured free-text clinical notes for patients. Wikipedia pages corresponding to the diagnoses in the MIMIC−III notes are utilized as our external knowledge source. Empirical results and analyses revealed that C-MemNN improves the accuracy of clinical diagnostic inferencing over other

classes of memory networks by a considerable margin (up to 23% improvement in average precision over the top five predictions with higher number of memory hops) [69].

The efficacy of a supervised machine learning model largely depends on the size of the annotated datasets used for training. Creation of labeled datasets requires expert-derived annotations, which are typically very expensive and time-intensive to obtain. To address the scarcity of large annotated datasets, we also formulated the diagnostic inferencing problem as a sequential decision making process using deep reinforcement learning [71].

Extracting appropriate clinical concepts from free clinical text is a critical first step for diagnosis inferencing. Existing clinical concept extraction tools are limited to the original content of the text as they do not consider evidence from external resources. Hence, clinical concepts extracted by these tools often lack aspects related to in-domain normalization, which may have a negative impact on the downstream clinical inferencing task. External (online) health-related resources can serve as the evidence to improve the original extracted concepts using one of the following ways: mapping of incomplete concepts to corresponding expressive concepts e.g. $personality \rightarrow personality\ changes$, paraphrasing the concepts e.g. *poor memory* $\rightarrow$ *memory loss*, and supplementing with additional concepts.

We proposed a novel clinical diagnosis inferencing approach that uses a deep reinforcement learning technique via a MDP formulation to incrementally learn about the most appropriate clinical concepts that best describe the correct diagnosis by using evidences gathered from relevant external resources (Figure 5). During training, the agent tries to learn the optimal policy through iterative search and consolidation of the most relevant clinical concepts related to the given patient condition. A deep Q-network architecture [26] is trained to optimize a reward function that measures the accuracy of the candidate diagnoses and clinical concepts. Our preliminary experiments on the Text REtrieval Conference (TREC) Clinical Decision Support (CDS) track[3] dataset [72] demonstrated the effectiveness of our system over various non-reinforcement learning-based baselines (up to 104% improvement in Mean Reciprocal Rank (MRR) scores and up to 56% improvement in average recall at the top 5 diagnoses) [71].

Recently, we proposed another novel approach for clinical diagnostic inferencing that focuses on the clinician's cognitive process to infer the most probable diagnoses from clinical narratives. Given a clinical text scenario, physicians typically review the sentences sequentially, skipping irrelevant parts and focusing on those that would contribute to the overall understanding of the clinical scenario. While assimilating the sentences, clinicians generally try to recognize a logical pattern or clinical progression similar to one or more prior patient encounters towards arriving at a provisional diagnosis. Ultimately, the intuition of the clinicians is guided by understanding these sentences and they can make an overall assessment of the scenario based on the narrative and/or additional evidence obtained from relevant external knowledge sources. Our new system replicated this cognitive flow by us-

[3] http://www.trec-cds.org/

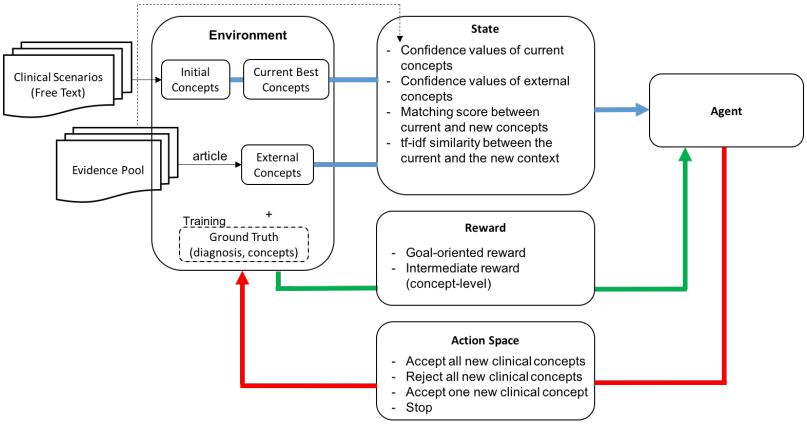

Fig. 5: Clinical diagnostic inferencing via improving clinical concept extraction with deep reinforcement learning [71].

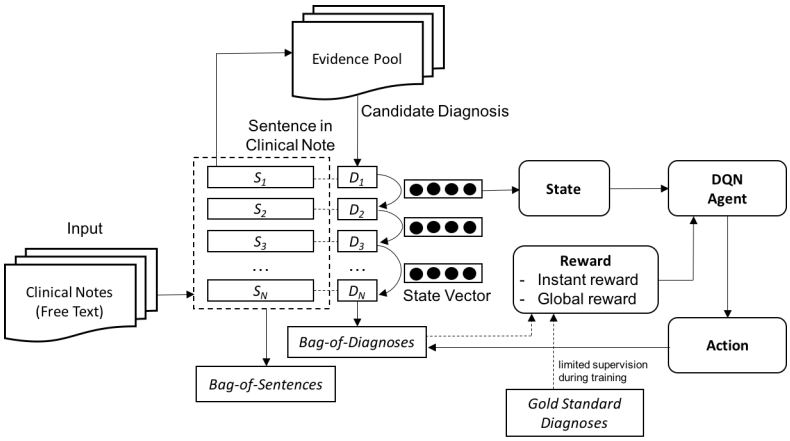

Fig. 6: Replicating clinician's cognitive process for clinical diagnostic inferencing using deep reinforcement learning [73].

ing a deep reinforcement learning technique (Figure 6). During training, the agent learns the optimal policy to obtain the final diagnoses through iterative search for candidate diagnoses from external knowledge sources via a sentence-by-sentence analysis of the inherent clinical context. A deep Q-network architecture [26] was trained to optimize a reward function that measures the accuracy of the candidate diagnoses. Our model predicted the differential diagnoses by utilizing the optimum policy learned to maximize the overall possible reward for an action during training. Extensive experiments on the TREC CDS track [72, 74] datasets demonstrated the effectiveness of this novel approach over several non-reinforcement learning-based systems (up to 100% improvement in terms of F-scores) [73].

We envisage that our recent works on clinical diagnostic inferencing can support the typically multitasking clinicians in considering some relevant differential diagnoses that could otherwise be ignored leading to inadvertent diagnostic errors. Also, relatively less skilled healthcare providers e.g. nurse practitioners can use the proposed system as a source of second opinion before contacting a physician towards accurately diagnosing and managing their patients.

3.2.2 Biomedical Article Retrieval

The main objective of the TREC CDS track was to retrieve a ranked list of the top biomedical articles that can answer generic clinical questions related to three categories: diagnosis, test, and treatment given a short clinical narrative.

We participated in this challenge [75] and our approach (Figure 7) centered on three steps: (i) Topical Keyword Analysis: identifying the most clinically relevant keywords from the given topic descriptions, summaries, and clinical notes using a clinical NLP engine [76]; (ii) Diagnostic Inferencing: reasoning based on the topical keywords to generate the diagnoses, tests, and treatments using the underlying clinical contexts represented within a Key-Value Memory Network, powered by an external clinical knowledge source; and, (iii) Relevant Article Retrieval: retrieving and ranking pertinent biomedical articles based on the topical keywords and clinical inferences from steps (i) and (ii).

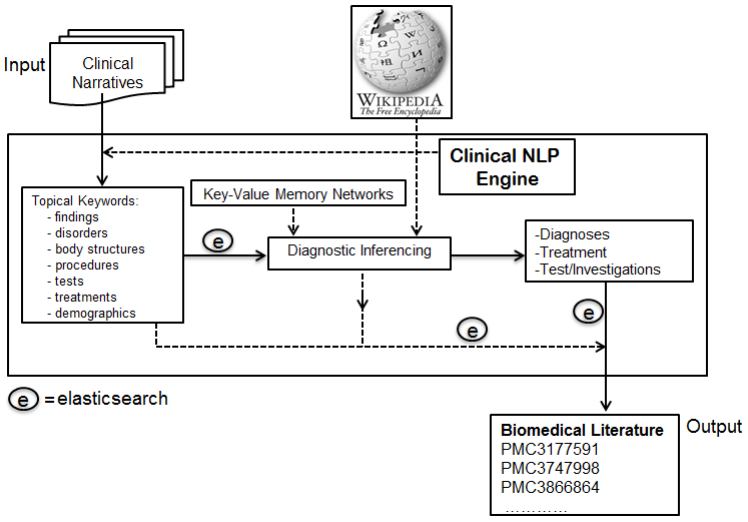

Fig. 7: System architecture for biomedical article retrieval.

We built a novel end-to-end diagnostic inferencing model using Key-Value Memory Networks [23] trained on a large collection of MIMIC-II discharge notes along

with the Wikipedia articles in the clinical medicine category in order to capture the overall context of a given clinical note towards inferring the most probable diagnoses. The list of possible diagnoses was then used to extract a list of candidate Wikipedia articles to mine related tests, and treatments (from sections and subsections of the Wikipedia article) accordingly. As the final step, topical keywords and the corresponding diagnoses, tests, and treatments obtained from the diagnostic inferencing step were used to retrieve candidate biomedical articles by searching through the given TREC-CDS corpus of over 1.25M PubMed Central[4] articles (indexed using Elasticsearch). Evaluation results showed additional gains with the use of the key-value memory network-based diagnostic inferencing approach for our clinical question answering system. In particular, on average our key-value memory network model with notes as input consistently outperformed the knowledge graph-based system for notes and descriptions as inputs in terms of infNDCG, R-prec, and Prec(10) scores. This system can be used to provide clinicians with biomedical articles containing scientific findings focused on a clinical scenario towards better-informed clinical decision-making.

3.2.3 Clinical Paraphrase Generation

Clinical paraphrase generation is important in building patient-centric decision support applications where users are able to understand complex clinical jargons via easily comprehensible alternative paraphrases. For example, the complex clinical term *"nocturnal enuresis"* can be paraphrased as *"nocturnal incontinence of urine"* or *"bedwetting"* to better clarify a well-known condition associated with children. We proposed *Neural Clinical Paraphrase Generation (NCPG)*, a novel approach to cast the clinical paraphrase generation task as a monolingual neural machine translation (NMT) problem. We used an end-to-end neural network in the form of an attention-based bidirectional RNN architecture within an encoder-decoder framework (Figure 8) to perform the task [60].

Extensive experiments on a large curated clinical paraphrase corpus built on a benchmark parallel paraphrase database, PPDB 2.0 [77], along with a comprehensive medical metathesaurus [78] show that the proposed attention-based NCPG model can outperform an RNN encoder-decoder based strong baseline for word-level modeling (up to 27% improvement in BLEU scores), whereas character-level models can achieve further improvements over their word-level counterparts (up to 25% improvement in BLEU scores). Table 1 shows a few example paraphrases generated by the proposed models.

Overall, the models demonstrate comparable performance relative to the state-of-the-art phrase-based conventional machine translation models (e.g. Moses). Recently, we further extended this work to go beyond lexical and phrasal paraphrasing, and proposed neural network-based models for sentence-level clinical paraphrase generation and simplification [79]. We believe that these models can be used to

[4] http://www.ncbi.nlm.nih.gov/pmc/

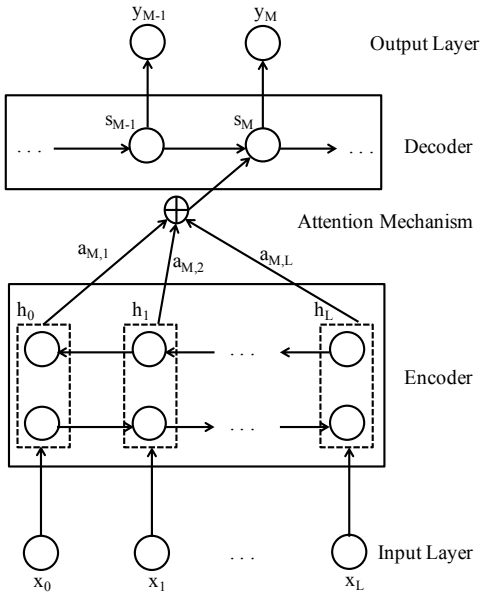

Fig. 8: Attention-based bidirectional RNN architecture for clinical paraphrase generation [60].

motivate patient engagement across the care continuum towards achieving desired outcomes.

Source: contagious diseases	**Target:** communicable diseases
Model	**Paraphrase**
Baseline (Word)	habitat
Baseline (Char)	contact diseases
NCPG (Word)	an infectious disease
NCPG (Char)	the diseases
Phrase-based Model (Moses)	infectious diseases
Source: secondary malignant neoplasm of spleen	**Target:** secondary malignant deposit to spleen
Model	**Paraphrase**
Baseline (Word)	secondary cancer of spleen
Baseline (Char)	separation of spleen
NCPG (Word)	secondary malignant neoplasm of spleen
NCPG (Char)	secondary malignant neoplasm
Phrase-based Model (Moses)	metastatic ca spleen

Table 1: Paraphrase examples [60].

3.2.4 Adverse Drug Event (ADE) Detection from Social Media

Adverse Drug Events (ADE) refer to negative side effects that may occur as a result of medication use. Monitoring and detection of such events (also called, *Pharmacovigilance*) is necessary to minimize potential health risks of patients by issuing warnings or recommending possible withdrawals of harmful pharmaceutical products.

Following pharmaceutical development, drugs are typically approved for use by the general public after going through clinical trials in limited settings. It is often impossible to uncover all adverse effects during these clinical trials. To address this issue, pharmaceutical and regulatory organizations require post-market surveillance programs to capture previously undiscovered adverse events. Traditional post-market ADE surveillance systems suffer from under-reporting and significant time delays in data processing, resulting in high incidence of unidentified adverse events related to medication use.

In the past decade, the rise of social media platforms (e.g. Twitter) has revolutionized online communication and networking. Due to the high *volume and velocity* of messages generated and distributed, social media data has been used for real-time information retrieval and trends tracking, including digital disease surveillance. Hence, we proposed a semi-supervised CNN-based architecture (Figure 9) that automatically detects ADEs as described in social media (e.g. Twitter feeds) [14].

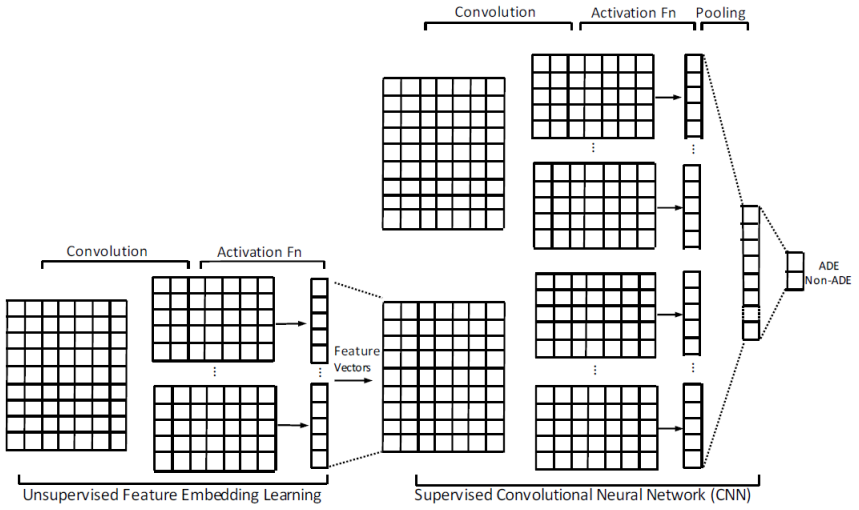

Fig. 9: Semi-supervised CNN architecture for ADE detection [14].

Unlike conventional systems that typically rely on large amounts of labeled data to train supervised machine learning models, our system can efficiently learn from

large volumes of unlabeled data in combination with a relatively small *seed set* of labeled ADEs. Our experimental results showed that the proposed system achieves better performance compared to traditional supervised machine learning algorithms for recommendations of ADEs from real-time social media streams (up to 9.9% improvement in F1-scores) [14]. The proposed system can be used to augment official post-market ADE surveillance systems. Readers are referred to [14] for additional technical details and analyses.

3.2.5 Medical Image Caption Generation

Visual perception and cues remain an important component for efficient understanding of natural language. Automatically understanding the content of an image and describing in natural language is a challenging task which has gained a lot of attention from computer vision and NLP researchers in recent years through various challenges for visual recognition and caption generation.

Due to the ever-increasing number of images in the medical domain that are generated across the clinical diagnostic pipeline, automated understanding of the image content could especially be beneficial for clinicians to provide useful insights and reduce their overall cognitive burden during patient care. Motivated by this need for automated image understanding methods in the healthcare domain, ImageCLEF[5] recently organized its inaugural caption prediction and concept detection tasks [80, 81]. The main objective of the concept detection task was to retrieve the relevant clinical concepts (e.g. anatomy, finding, diagnosis) that are reflected in a medical image. Whereas in the caption prediction task, participants were supposed to leverage the clinical concept vocabulary created in the concept detection task towards generating a coherent caption for each medical image.

We submitted several runs for caption prediction and concept detection tasks by using an attention-based image caption generation framework (Figure 10). The attention mechanism automatically learns to emphasize on salient parts of the medical image while generating corresponding words in the output for the caption prediction task and corresponding clinical concepts for the concept detection task. In particular, motivated by the success of prior works in solving general domain image captioning tasks, we used an encoder-decoder based deep neural network architecture for the caption prediction task [84], where the encoder uses a deep CNN [85] to encode a raw medical image to a feature representation, which is in turn decoded using an attention-based RNN to generate the most relevant caption for the given image. Figure 11 shows an example caption generated by our proposed model.

We followed a similar approach to address the concept detection task by treating it as a text generation problem. Our system was ranked first (with mean BLEU score of 0.32) in the caption prediction task among submissions with no prior exposure to the test set, while we showed a decent performance (with mean F1-score of 0.12) in

[5] http://www.imageclef.org/2017/caption

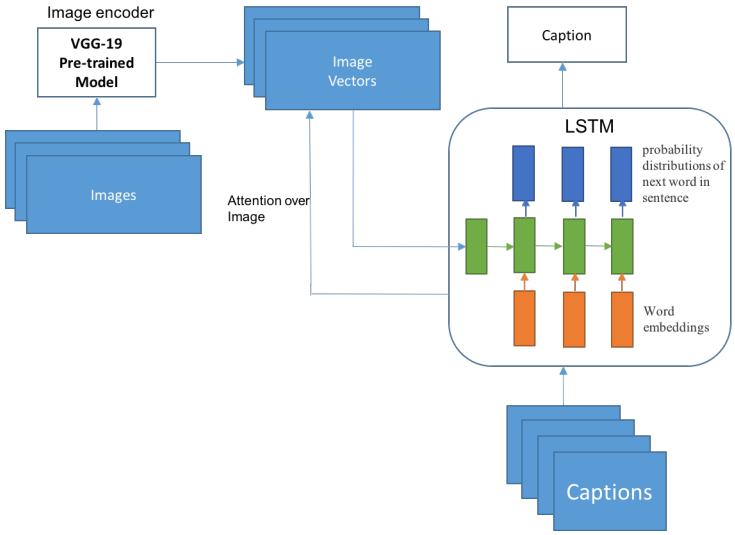

Fig. 10: Attention-based image caption generation framework [82, 83].

Ground Truth:	CT scan of the abdomen with contrast of Case 2 showing a large, loculated liver abscess measuring 10 cm.
Model:	ct scan of the abdomen on the first visit shows an irregular huge low density mass .

Fig. 11: Example caption generated by our model.

the concept detection task. Interested readers are referred to [82, 83] for additional details and examples.

4 Conclusion

In this tutorial chapter, we have presented an overview of how deep learning techniques can be applied to solve NLP tasks in general, followed by a literature survey of existing deep learning algorithms applied to clinical NLP problems, and finally, a description of various deep learning-driven clinical NLP applications developed at the Artificial Intelligence (AI) lab in Philips Research in recent years - such as diagnostic inferencing from unstructured clinical narratives, relevant biomedical article retrieval based on clinical case scenarios, clinical paraphrase generation, adverse drug event (ADE) detection from social media, and medical image caption generation. Our proposed models have demonstrated the effectiveness of deep learning techniques to address various clinical NLP problems as they achieved state-of-the-art results compared to lexicon-, knowledge source-, and traditional machine learning-based systems.

References

1. Mona Alsaffar, Peter Yellowlees, Alberto Odor, and Michael Hogarth. The state of open source electronic health record projects: A software anthropology study. *JMIR Medical Informatics*, 5(1):e6, 2017.
2. Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.
3. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.
4. Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.
5. Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015*, pages 1556–1566, 2015.
6. G. W. Stewart. On the early history of the singular value decomposition. *SIAM Review*, 35(4):551–566, 1993.
7. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.
8. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems NIPS 2013*, pages 3111–3119, 2013.
9. Thang Luong, Richard Socher, and Christopher D. Manning. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning, CoNLL 2013*, pages 104–113, 2013.
10. Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1188–1196, 2014.
11. Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, 2014.

12. Yoav Goldberg. *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers, 2017.
13. R. Johnson and T. Zhang. Effective use of word order for text categorization with convolutional neural networks. In *Proceedings of the 11th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL–HLT)*, pages 103–112, Denver, Colorado, 2015.
14. Kathy Lee, Ashequl Qadir, Sadid A. Hasan, Vivek V. Datla, Aaditya Prakash, Joey Liu, and Oladimeji Farri. Adverse drug event detection in tweets with semi-supervised convolutional neural networks. In *Proceedings of the 26th International Conference on World Wide Web, WWW*, pages 705–714, 2017.
15. Jonas Gehring, Michael Auli, David Grangier, and Yann N Dauphin. A Convolutional Encoder Model for Neural Machine Translation. *ArXiv e-prints*, 2016.
16. Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional Sequence to Sequence Learning. *ArXiv e-prints*, 2017.
17. I. Sutskever, J. Martens, and G. E. Hinton. Generating Text with Recurrent Neural Networks. In *Proceedings of ICML*, pages 1017–1024, 2011.
18. Y. Bengio, P. Simard, and P. Frasconi. Learning Long-Term Dependencies with Gradient Descent is Difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.
19. S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.
20. K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, 2014.
21. Sainbayar Sukhbaatar, arthur szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *NIPS*, pages 2440–2448, 2015.
22. Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *CoRR*, abs/1410.3916, 2014.
23. Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. *CoRR*, abs/1606.03126, 2016.
24. Christopher J. C. H. Watkins and Peter Dayan. Q-learning. In *Machine Learning*, pages 279–292, 1992.
25. Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction* . The MIT Press, Cambridge, Massachusetts, London, England, 1998.
26. Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.
27. Simon Baker, Anna Korhonen, and Sampo Pyysalo. Cancer hallmark text classification using convolutional neural networks. In *(BioTxtM)*, pages 1–9, 2016.
28. Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. Representations of time expressions for temporal relation extraction with convolutional neural networks. In *BioNLP*, pages 322–327, 2017.
29. Sunil Mohan, Nicolas Fiorini, Sun Kim, and Zhiyong Lu. Deep learning for biomedical information retrieval: Learning textual relevance from click logs. In *BioNLP*, pages 222–231, 2017.
30. Yifan Peng and Zhiyong Lu. Deep learning for extracting protein-protein interactions from biomedical literature. In *BioNLP*, pages 29–38, 2017.
31. Masaki Asada, Makoto Miwa, and Yutaka Sasaki. Extracting drug-drug interactions with attention cnns. In *BioNLP*, pages 9–18, 2017.
32. Matthew C Chen, Robyn L Ball, Lingyao Yang, Nathaniel Moradzadeh, Brian E Chapman, David B Larson, Curtis Langlotz, Timothy J Amrhein, and Matthew Lungren. Deep Learning to Classify Radiology Free-Text Reports. *Radiology*, 2017.

33. Lina Sulieman, David Gilmore, Christi French, Robert M. Cronin, Gretchen Purcell Jackson, Matthew Russell, and Daniel Fabbri. Classifying patient portal messages using convolutional neural networks. *Journal of Biomedical Informatics*, 74:59–70, 2017.
34. Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. A neural network multi-task learning approach to biomedical named entity recognition. *BMC Bioinformatics*, 18(1), 2017.
35. Sarvnaz Karimi, Xiang Dai, Hamedh Hassanzadeh, and Anthony Nguyen. Automatic diagnosis coding of radiology reports: A comparison of deep learning and conventional classification methods. In *BioNLP*, pages 328–332, 2017.
36. R. Feldman, O. Netzer, A. Peretz, and B. Rosenfeld. Utilizing text mining on online medical forums to predict label change due to adverse drug reactions. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1779–1788, Sydney, Australia, 2015.
37. J. Lardon, R. Abdellaoui, F. Bellet, H. Asfari, J. Souvignet, N. Texier, M. C. Jaulent, M. N. Beyens, A. Burgun, and C. Bousquet. Adverse drug reaction identification and extraction in social media: A scoping review. *Journal of Medical Internet Research*, 17(7):e171, 2015.
38. A. Sarker, R. E. Ginn, A. Nikfarjam, K. O'Connor, K. Smith, S. Jayaraman, T. Upadhaya, and G. Gonzalez. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202–212, 2015.
39. M. Yang, M. Kiang, and W. Shang. Filtering big data from social media - building an early warning system for adverse drug reactions. *Journal of Biomedical Informatics*, 54(C):230–240, 2015.
40. A. Sarker and G. Gonzalez. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics*, 53:196 – 207, 2015.
41. X. Liu and H. Chen. Identifying adverse drug events from patient social media: A case study for diabetes. *IEEE Intelligent Systems*, 30(3):44–51, 2015.
42. Abhyuday N Jagannatha and Hong Yu. Bidirectional rnn for medical event detection in electronic health records. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 473–482, 2016.
43. Abhyuday Jagannatha and Hong Yu. Structured prediction models for RNN based sequence labeling in clinical text. In *EMNLP*, pages 856–865, 2016.
44. Adyasha Maharana and Meliha Yetisgen. Clinical Event Detection with Hybrid Neural Architecture. In *BioNLP*, pages 351–355, 2017.
45. Shweta Yadav, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. Deep learning architecture for patient data de-identification in clinical records. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 32–41, 2016.
46. Franck Dernoncourt, Ji Young Lee, Özlem Uzuner, and Peter Szolovits. De-identification of patient notes with recurrent neural networks. *JAMIA*, 24(3):596–606, 2017.
47. Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. De-identification of Clinical Notes via Recurrent Neural Network and Conditional Random Field. *Journal of Biomedical Informatics*, 75, 2017.
48. Wael Salloum, Greg Finley, Erik Edwards, Mark Miller, and David Suendermann-Oeft. Deep learning for punctuation restoration in medical reports. In *BioNLP*, pages 159–164, 2017.
49. Rahul V S S Patchigolla, Sunil Sahu, and Ashish Anand. Biomedical event trigger identification using bidirectional recurrent neural network based models. In *BioNLP*, pages 316–321, 2017.
50. Hua He, Kris Ganjam, Navendu Jain, Jessica Lundin, Ryen White, and Jimmy Lin. An insight extraction system on biomedical literature with deep neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2691–2701, 2017.
51. Raghavendra Chalapathy, Ehsan Zare Borzeshi, and Massimo Piccardi. Bidirectional LSTM-CRF for clinical concept extraction. In *ClinicalNLP@COLING 2016*, pages 7–12, 2016.

52. Inigo Jauregi Unanue, Ehsan Zare Borzeshi, and Massimo Piccardi. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. *Journal of Biomedical Informatics*, 76:102–109, 2017.
53. Zengjian Liu, Ming Yang, Xiaolong Wang, Qingcai Chen, Buzhou Tang, Zhe Wang, and Hua Xu. Entity recognition from clinical texts via recurrent neural network. *BMC Medical Informatics and Decision Making*, 17(2), 2017.
54. Gabriel Stanovsky, Daniel Gruhl, and Pablo Mendes. Recognizing mentions of adverse drug reaction in social media using knowledge-infused recurrent models. In *EACL*, pages 142–151, 2017.
55. Sunil Kumar Sahu and Ashish Anand. Recurrent neural network models for disease name recognition using domain invariant features. In *ACL*, 2016.
56. N. Elhadad and K. Sutaria. Mining a Lexicon of Technical Terms and Lay Equivalents. In *Proceedings of the Workshop on BioNLP*, pages 49–56, 2007.
57. L. Deléger and P. Zweigenbaum. Extracting Lay Paraphrases of Specialized Expressions from Monolingual Comparable Medical Corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: From Parallel to Non-parallel Corpora*, pages 2–10, 2009.
58. C. Wang, L. Cao, and B. Zhou. Medical Synonym Extraction with Concept Space Models. In *Proceedings of IJCAI*, pages 989–995, 2015.
59. Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. Neural paraphrase generation with stacked residual lstm networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2923–2934, 2016.
60. Sadid A. Hasan, Bo Liu, Joey Liu, Ashequl Qadir, Kathy Lee, Vivek V. Datla, Aaditya Prakash, and Oladimeji Farri. Neural clinical paraphrase generation with attention. In *Proceedings of the Clinical Natural Language Processing Workshop, ClinicalNLP@COLING*, pages 42–53, 2016.
61. Edward Choi, Mohammad Taha Bahadori, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. *arXiv preprint arXiv:1511.05942*, 2015.
62. Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Retain: Interpretable predictive model in healthcare using reverse time attention mechanism. *CoRR*, abs/1608.05745, 2016.
63. Stelmo Magalhães Barros Netto, Anselmo Cardoso de Paiva, Areolino de Almeida Neto, Aristofanes Correa Silva, and Vanessa Rodrigues Coelho Leite. *Application on Reinforcement Learning for Diagnosis Based on Medical Image*. INTECH Open Access Publisher, 2008.
64. Radhika Poolla. A reinforcement learning approach to obtain treatment strategies in sequential medical decision problems. *Graduate Theses and Dissertations, University of South Florida*, 2003.
65. Susan M Shortreed, Eric Laber, Daniel J Lizotte, T Scott Stroup, Joelle Pineau, and Susan A Murphy. Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Machine learning*, 84(1-2):109–136, 2011.
66. Yufan Zhao, Donglin Zeng, Mark A Socinski, and Michael R Kosorok. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67(4):1422–1433, 2011.
67. Karthik Narasimhan, Tejas D. Kulkarni, and Regina Barzilay. Language understanding for text-based games using deep reinforcement learning. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1–11, 2015.
68. Karthik Narasimhan, Adam Yala, and Regina Barzilay. Improving information extraction by acquiring external evidence with reinforcement learning. *arXiv preprint arXiv:1603.07954*, 2016.
69. Aaditya Prakash, Siyuan Zhao, Sadid A. Hasan, Vivek V. Datla, Kathy Lee, Ashequl Qadir, Joey Liu, and Oladimeji Farri. Condensed memory networks for clinical diagnostic inferencing. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 3274–3280, 2017.

70. Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3, 2016.
71. Yuan Ling, Sadid A. Hasan, Vivek Datla, Ashequl Qadir, Kathy Lee, Joey Liu, and Oladimeji Farri. Diagnostic inferencing via improving clinical concept extraction with deep reinforcement learning: A preliminary study. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, pages 271–285, 2017.
72. Kirk Roberts, Matthew S Simpson, Ellen Voorhees, and William R Hersh. Overview of the trec 2015 clinical decision support track. In *TREC*, 2015.
73. Yuan Ling, Sadid A. Hasan, Vivek Varma Datla, Ashequl Qadir, Kathy Lee, Joey Liu, and Oladimeji Farri. Learning to diagnose: Assimilating clinical narratives using deep reinforcement learning. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP*, pages 895–905, 2017.
74. Kirk Roberts, Dina Demner-Fushman, Ellen Voorhees, and William R Hersh. Overview of the TREC 2016 Clinical Decision Support Track. In *TREC*, 2016.
75. Sadid A Hasan, Siyuan Zhao, Vivek Datla, Joey Liu, Kathy Lee, Ashequl Qadir, Aaditya Prakash, and Oladimeji Farri. Clinical question answering using key-value memory networks and knowledge graph. In *TREC*, 2016.
76. V. Datla, S. A. Hasan, A. Qadir, K. Lee, Y. Ling, J. Liu, and O. Farri. Automated Clinical Diagnosis: The Role of Content in Various Sections of a Clinical Document. In *IEEE-BIBM International Workshop on Biomedical and Health Informatics*, 2017.
77. E. Pavlick, P. Rastogi, J. Ganitkevitch, B. Van Durme, and C. Callison-Burch. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of ACL-IJCNLP)*, pages 425–430, 2015.
78. D. Lindberg, B. Humphreys, and A. McCray. The Unified Medical Language System. *Methods of Information in Medicine*, 32(4):281–291, 1993.
79. Viraj Adduru, Sadid A. Hasan, Joey Liu, Yuan Ling, Vivek Datla, Kathy Lee, Ashequl Qadir, and Oladimeji Farri. Towards Dataset Creation and Establishing Baselines for Sentence-level Neural Clinical Paraphrase Generation and Simplification. In *Proceedings of the 3rd International Workshop on Knowledge Discovery in Healthcare Data (KDH) @ IJCAI-ECAI*, 2018.
80. Bogdan Ionescu, Henning Müller, Mauricio Villegas, Helbert Arenas, Giulia Boato, Duc-Tien Dang-Nguyen, Yashin Dicente Cid, Carsten Eickhoff, Alba Garcia Seco de Herrera, Cathal Gurrin, Md Bayzidul Islam, Vassili Kovalev, Vitali Liauchuk, Josiane Mothe, Luca Piras, Michael Riegler, and Immanuel Schwall. Overview of imageclef 2017: Information extraction from images. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 8th International Conference of the CLEF Association, CLEF Proceedings*, pages 315–337, 2017.
81. Carsten Eickhoff, Immanuel Schwall, Alba Garcia Seco de Herrera, and Henning Müller. Overview of imageclefcaption 2017 - image caption prediction and concept detection for biomedical images. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum*, 2017.
82. Sadid A. Hasan, Yuan Ling, Joey Liu, Rithesh Sreenivasan, Shreya Anand, Tilak Raj Arora, Vivek V. Datla, Kathy Lee, Ashequl Qadir, Christine Swisher, and Oladimeji Farri. PRNA at imageclef 2017 caption prediction and concept detection tasks. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum*, 2017.
83. Sadid A. Hasan, Yuan Ling, Joey Liu, Rithesh Sreenivasan, Shreya Anand, Tilak Raj Arora, Vivek Datla, Kathy Lee, Ashequl Qadir, Christine Swisher, and Oladimeji Farri. Attention-based Medical Caption Generation with Image Modality Classification and Clinical Concept Mapping. In *Proceedings of the 9th International Conference and Labs of the Evaluation Forum (CLEF)*, 2018.
84. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 2048–2057, 2015.
85. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.