Neural Natural Language Processing for Unstructured Data in Electronic Health Records: a Review

IRENE LI, JEREMY GOLDWASSER, JESSICA PAN, WAI PAN WONG, YAVUZ NUZUMLALI, NEHA VERMA, BENJAMIN ROSAND, YIXIN LI, MATTHEW ZHANG, DAVID CHANG, R. ANDREW TAYLOR, DRAGOMIR RADEV, Yale University

Electronic health records (EHRs), digital collections of patient healthcare events, are ubiquitous in medicine and critical to healthcare delivery, operations, and research. Despite this central role, well over half of the information stored within EHRs is in the form of unstructured text (e.g. provider notes, operation reports) and remains largely untapped for secondary use. Traditional methods to leverage this information via natural language processing (NLP) techniques for EHRs have been hampered by XX, XX, XX. Recently, however, newer neural network and deep learning approaches in NLP have made considerable recent advances outperforming traditional statistical and rule-based systems on a variety of tasks. In this survey paper, we summarize current state-of-the-art neural NLP methods for EHR applications. We focus on a broad scope of tasks, namely, classification and prediction, clinical embeddings, information extraction, generation, summarization and simplification, and other topics including question answering, phenotyping, knowledge graphs, multilinguality and medical dialogues.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Computing methodologies** → **Natural language processing**; **Machine learning algorithms**.

Additional Key Words and Phrases: natural language processing, neural networks, EHR, unstructured data

ACM Reference Format:
Irene Li, Jeremy Goldwasser, Jessica Pan, Wai Pan Wong, Yavuz Nuzumlali, Neha Verma, Benjamin Rosand, Yixin Li, Matthew Zhang, David Chang, R. Andrew Taylor, Dragomir Radev. 2018. Neural Natural Language Processing for Unstructured Data in Electronic Health Records: a Review. *Proc. ACM Meas. Anal. Comput. Syst.* 37, 4, Article 111 (August 2018), 47 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION OR BACKGROUND

Electronic health records (EHRs), digital collections of patient healthcare events, are now ubiquitous in medicine and critical to healthcare delivery, operations, and research. [64]. [39, 42, 56, 82] Data within EHRs are often classified based on collection and representation formats as belonging to one of two classes: structured or unstructured. [37] Structured EHR data consist of heterogeneous sources like diagnoses, medications, and laboratory values in fixed numerical or categorical fields; unstructured data, in contrast, refers to free-form text written by healthcare providers, such as clinical notes and discharge summaries. Unfortunately, unstructured data represents about 80% of EHR data and remains largely untapped for secondary use. [247] In this survey paper, we focus our discussion on unstructured text data in the EHR and newer neural, deep-learning based methods employed to leverage this type of data.

Author's address: Irene Li, Jeremy Goldwasser, Jessica Pan, Wai Pan Wong, Yavuz Nuzumlali, Neha Verma, Benjamin Rosand, Yixin Li, Matthew Zhang, David Chang, R. Andrew Taylor, Dragomir RadevYale University.

2476-1249/2018/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

[25, 156] In recent years, artificial neural networks have dramatically impacted fields such as speech recognition, computer vision (CV), and natural language processing (NLP) within medicine and elsewhere. These models, the defining aspect of deep learning, have come to outperform and supersede traditional, rule-based methods on many tasks. The application of neural text mining methods to unstructured data in EHRs using NLP is attracting great interest among researchers in artificial intelligence (AI). Natural language processing (NLP) is the field concerned with using computers to process human, or "natural," language. More recently, neural methods inspired by the human brain have come to outperform statistical and rule-based systems in NLP.[61] In order to successfully train, neural networks require huge numbers of training examples and model parameters; these requirements prevented deep learning from coming into favor for a variety of purposes until the late 2000s, when the technology for efficient computation and vast data storage finally became available.

1.1 Challenges and difficulties

In this paper we focus on neural approaches to analyzing unstructured data. Unstructured EHR data can carry abundant useful information in healthcare, but it is difficult to analyze because of a number of challenges and difficulties.

Privacy Unlike in other domains, privacy is an important issue when researchers work with EHR data. [54, 105] So before any downstream tasks can be done, additional manuevers are required for ensuring patient privacy. Removing identifying information from a large corpus of EHRs is an expensive process, as it difficult to automate and usually requires annotators with strong domain expertise. For this reason, one of the sections discussed in this paper is automatic de-identification of patient records.

Lack of annotations Many machine learning and deep learning models are supervised models and thus labeled data is necessary for training. Annotating EHR data can be challenging based on workloads and data quality. Besides, for some certain tasks, only qualified annotators can be recruited to complete annotations. Even though annotations can be done, the quality of the annotations are sometimes hard to be ensured; there may still be disagreements between annotators, making evaluations more difficult and controversial. Therefore, useful EHR data for training is often in short supply.

Interpretability While deep neural networks are able to achieve superior results compared with other methods in many fields, they are often treated as a black-box. [222, 264] Typically, a neural network model has a large number of trainable parameters, and as brought difficulties for model interpretability. Moreover, unlike linear models, which are usually more straightforward and explainable, neural networks consist of non-linear layers which are not quite interpretable. Recently, there have been a few attempts to produce explainable deep neural networks to increase model transparency. [23, 24, 153]

1.2 Related Work

Prior surveys have focused on a variety of deep learning topics within health informatics, bioinformatics, EHRs with Neural NLP methods. An early survey by Miotto et al [145] summarized deep learning methods for healthcare and applications in clinical imaging, EHRs, Genomics and Mobile domains. Various works were reviewed in EHRs for predicting diseases, modeling phenotypes, and learning representations of medical concepts, such as diseases and medications. Some other survey papers [3, 194] described clinical applications utilizing deep learning techniques including: Information Extraction, Representation Learning, Outcome Prediction, Phenotyping and De-identification. Similarly, in a systematic survey[247], five categories were targeted: disease detection/classification, sequential prediction, concept embedding, data augmentation and EHR data privacy. Kwak and Hui[106] reviewed breakthroughs of research applying artificial intelligence in health informatics. Especially, in the EHR field, they included the following applications: Outcome Prediction, Computational Phenotyping, Knowledge Extraction, Representation Learning, De-identification and Medical Intervention Recommendations. [6] presents a literature

review of how free-text content electronic patient records could benefit from recent Natural Language Processing techniques, by selecting four application domains: sentiment analysis and predictive models, and automated patient cohort selection. Another recent survey [243] summarized deep learning methods in clinical NLP, and the authors reviewed methods such as deep learning architectures, embeddings and medical knowledge on four groups of clinical NLP tasks: text classification, named entity recognition, relation extraction, and others. [89] discusses textual epidemic intelligence, or the detection of disease outbreaks via medical and informal (i.e. the Web) text. This survey includes approaches via NLP to query and analyze medical records, as well as other data sources, for indications of burgeoning epidemics.

1.3 Summary of review focus

While the mentioned reviews have different targeting applications or techniques, and varied on the scope of selected topics, in our survey, we look at a more comprehensive coverage of applications and tasks, and we include typical works with very recent BERT-based models. Specifically, we summarize a broad range of existing literature on deep learning methods with EHR data. Most of the papers on clinical NLP discussed achieve performance near the state-of-the-art for their task; for this reason, we limit our scope on the majority of these works to papers published after 2015. We also reference several older papers, both in and out of the clinical domain, that had a strong influence on subsequent works.

In this survey, we first provide some NLP preliminaries, then we explore the following main EHR and NLP tasks: classification and prediction, clinical embeddings, extraction, generation and summarization, and other topics including question answering, phenotyping, knowledge graphs, multilinguality and medical Dialogues. Finally, we also summarize relevant datasets and existing tools in the end.

2 PRELIMINARIES

In this section, we present an overview of important concepts in natural language processing and deep learning.

2.1 Natural Language Processing

The challenge of getting machines to understand human language dates back over a half century. In the 1950s, computer scientist Alan Turing formulated the famous idea of the Turing Test[216], which gauges machine intelligence based on its ability to mimic natural language. A computer passes the test if a person is unable to distinguish whether the responder is a machine or a human. No computer at the time was remotely capable of passing the Turing Test, of course. But it seeded the idea of natural language processing being a key form of artificial intelligence.

In the decades following the Turing Test, most NLP systems used rules to perform tasks like machine translation.[87] As one might imagine, these rule-based systems never achieved stellar performance because natural language is simply too complex to model by such rules. Starting in the late 1980s, researchers began shifting away from rule-based NLP and towards statistical methods. These methods cast language in a statistical and probabilistic framework, learning rules implicitly rather than naming them explicitly. Statistical NLP algorithms were able to take advantage of the explosion of text data available from the advent of the internet.

2.2 Machine Learning

Supervised learning is the task of learning a function that maps an input to an output.[148, 183] A supervised learning model would train to input images of cats and dogs and classify them to a high degree of accuracy. Supervised models require labeled datasets of example inputs and their corresponding outputs. Such labeled datasets are often unavailable or small. **Unsupervised learning** [148] models circumvent this problem, as they don't need labeled data to train. Rather, they train directly on the raw data itself. This makes them extremely

useful for NLP, where enormous free-text datasets like the entirety of English Wikipedia are in no short supply. Unsupervised neural methods are particularly useful, as deep learning models require very large amounts of data to train.

Representation Learning The performance of machine learning models is strongly dependent upon feature inputs and their representation. Historically, these features were selected and/or engineered via manual processes that utilized domain knowledge. This form of manual feature engineering can be very inefficient, as expert domain knowledge is often unavailable or even incorrect. Moreover, it frequently fails to account for the complexity inputs like images or natural language text. Representation learning is the subset of machine learning in which models automatically discern the important feature representations from raw data and avoids the problems of these rule-based systems by learning features implicitly.[189, 270] A key success of newer deep learning methods in NLP is their enhanced ability in the domain of representational learning.

Reinforcement learning Reinforcement learning (RL) is the task of learning to make a sequence of decisions that optimize long-term reward. [62] In an RL framework, an agent makes decisions that respond to, and affect, its environment. The field of robotics offers a classic example: a reinforcement learning algorithm would train a robot to learn to walk, rewarding it based on how far it travels and how long it remains upright. Reinforcement learning is less common in NLP, but it is used by some of the papers discussed in this survey.

2.3 Neural Networks

The human brain is a web of billions of interconnected neurons. [73] While as a whole it enables us to do enormously complicated things like see, speak, run, read, and remember, the individual neurons that constitute it are rather simple. Neurons are able to pass electrical or chemical signals to one another via the synapse, a small gap that connects two cells. Signals from multiple input channels enter a neuron through its dendrites; if the total signal received is above a certain threshold, the neuron will fire an action potential by rapidly depolarizing and polarizing again. It then transmits this output spike through the axon terminal, passing it as input to other neuron cells. These straightforward binary threshold neurons join en masse to form an organ arguably more powerful than the biggest supercomputer on Earth. [62, 73]

Artificial neural networks (ANNs), commonly referred to as neural networks, are models that emulate this system.[189] They use the same general architecture: layer after layer of neuron-like units that process the outputs of the previous layer. When a neural network has many layers, it is said to be "deep" - hence the name of the field that studies them. In the most basic of these architectures, a *feedforward neural network* [71], each neuron takes inputs from all neurons in the previous layer and passes its output to all neurons in the subsequent one. These layers gradually build up information that implicitly represents the relevant "features" of its initial input. This information can be used to perform tasks such as image classification or machine translation.

In order to perform these tasks, though, neural networks must learn the weight and bias parameters with which artificial neurons process their inputs. [81, 189] Typically, each unit takes a weighted sum of its inputs, adds a bias constant, and passes this to a predetermined activation function. Activation functions [157] like the Rectified Linear Unit (ReLU) are more expressive than the binary threshold output used by our brain's neurons. To train the network, we pass training examples through our model then estimate each parameter's contribution to the overall model error. These contributions are the loss gradients, which are calculated with the backpropagation algorithm. [182] We make small adjustments to the parameters according to these gradient values and repeat until the model converges.

2.4 Autoencoders

Autoencoders are a special kind of encoding mechanism.[101] The purpose of an encoder is to represent an input in a different number of dimensions (typically fewer) while preserving its salient features. In the autoencoder

framework, a neural network encodes an input, typically to a lower-dimensional representation space, and then decodes this encoding to reconstruct the original input. That way, the model learns to represent the input in such a way that its key properties can still be inferred.

There are several commonly-used twists on the autoencoder framework. The denoising autoencoder[224] corrupts each input by randomly turning some of its values to zero, but still trains to reconstruct the uncorrupted sample. Doing so prevents overfitting and forces the model to more robustly learn the dependencies between input features.

Another modification, the variational autoencoder[98], learns to represent each sample with a probability distribution rather than of a fixed encoding. This increases model robustness by accounting for greater variance in the latent space. And because a VAE learns a latent state distribution, its decoder can also generate new samples by randomly sampling encodings from the representation space.

A third variation is the stacked autoencoder.[225] This consists of not a single autoencoder, but rather a chain of successive ones. Each autoencoder within a stacked autoencoder inputs and reconstructs the encoding layer from the previous autoencoder. Stacked denoising autoencoders are particularly common in practice.

2.5 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) [57, 110, 230] are a classic neural model based on the human visual cortex. Each layer of a CNN convolves input matrices with smaller filters whose parameters are learned by the model. In effect, it will learn higher- and higher-level features with each successive layer. In the image processing case, it may detect edges in the first layers, piece together those low-level shapes in subsequent layers, and so on until it can identify abstract forms like faces or cats. For classification tasks, high-level CNN layers can be flattened into a vector that gets input to feedforward layers. CNNs are also applied for many NLP tasks such as text classification. [96, 265]

2.6 Recurrent Neural Networks

Recurrent Neural Networks (RNNs)[176, 241] process sequential input one step at a time. Because they share the same weights at every step, they are able to handle variable-length inputs; this useful property makes RNNs a common choice for NLP, which has textual input. [73, 152, 240] In the standard (or "Vanilla ") RNN, each time-step has a cell that processes both the input at that step as well as the processed input from all preceding steps. From these inputs, each cell outputs a hidden state that gets passed to the next RNN cell.

Vanilla RNNs struggle to learn long-term temporal dependencies because their gradients grow or decay exponentially over time-steps. To solve this issue, the Vanilla RNN cell can be replaced by an LSTM [70] or GRU cell [29]. These cells have gates that control the flow of information, making the network better at retaining memory over multiple time-steps.

2.7 Word Embeddings

Word embeddings map discrete tokens like words into a real-valued vector space, taking the word's semantics into account. Before the advent of deep learning, researchers typically represented words as one-hot vectors, with the 1 at the index corresponding to the word. This naive method had two major problems. Firstly, it completely ignored semantics; "Paris" would be just as close to "armchair" in representation space as it would to "France." Secondly, it was not scalable, as it represented each word as a vector over the entire vocabulary.

A seminal paper that used neural networks to create dense semantic word embeddings was Word2Vec. [142] Word2Vec introduced the Skip-Gram model, which learned word representations by predicting the surrounding context of a center word. It also introduced the Continuous Bag of Words (CBOW) model, which predicted a center word given its context. Both produced meaningful word embeddings that proved excellent at making

analogies. Later, Facebook AI created FastText [90], which modified Word2Vec by inputting n-gram character strings rather than whole words. This minor adjustment significantly sped up training. Like Word2Vec, GloVe [164] also makes embeddings based on the words that appear within a context window. GloVe, however, learns embeddings from global statistical information on the number of times words co-occur together.

Word2Vec, GloVe, and FastText embeddings produce fixed embeddings for a given word. This framework is limiting, however, because words may have many different meanings depending on their context. More recent models use RNNs to create such contextualized embeddings. Most of these are language models - models that predict the probability of an input sequence by factoring it as product of conditional probabilities. Each conditional probability is the probability of a word given the entire sequence that precedes it. Language models make excellent embedding models because they learn to represent words based on prior context.

An important model that used a language model to create word embeddings was ELMo [168]. ELMo, which stands for Embeddings from Language Models, made the novel development of using bidirectional LSTMs to consider a word's context in both directions. That way, each embedding factors in every other word in the input, not just the preceding ones.

2.8 Sequence to sequence models

As mentioned earlier, cells in an RNN can produce an output in addition to a hidden state. This gives them a large amount of flexibility: they can convert one input to one output, or one input to many outputs, or many inputs to one output, or many inputs to many outputs. Sequence to sequence (seq2seq) models are a special kind of many-to-many RNN, in which the whole input sequence is encoded before an output sequence is decoded one token at a time.[209] This presents an encoder-decoder framework resembling that of the autoencoder, but without the objective for the decoder to output the original input. At each decoder time-step, the seq2seq model generates a token in the output sequence; this gets passed as the input token for the next time-step. This repeats until the <EOS> (end-of-sequence) token is decoded.

2.9 Transformers

The advent of Transformer-based models [221] has been heralded as a breakthrough moment in NLP. The Transformer, designed by Google Brain researchers in 2017, is an encoder-decoder model that uses a modified attention mechanism called "self-attention" to represent text in a more parallelized fashion. Its novel network architecture trains faster and yields better results than traditional seq2seq models with attention.

Rather than using a sequential RNN to encode the input one token after another, the Transformer processes each input token at the same time. This parallelization dramatically speeds up training. But it does not process in isolation; instead, the self-attention mechanism examines the representations of all other input tokens when encoding an input. The overall architecture is a stack of encoder layers, the final result of which gets passed to a stack of decoder layers. As with standard seq2seq models, the Transformer decodes the output one token at a time.

Several models based on the Transformer have quite literally transformed the NLP landscape in the years following its release. First, OpenAI's Generative Pretrained Transformer (GPT) [171] was the first model to create word embeddings with Transformers. GPT, along with its subsequent versions GPT-2 and GPT-3, is a stack of Transformer decoders. All three are language models, as the decoder outputs one successive token given all previous decoded tokens at each timestep. With this framework, they are able to pretrain on vast amounts of unstructured text. Then, once pretrained, the GPT models can be adapted for a variety of tasks such as question answering and summarization.

Bidirectional Encoder Representations from Transformers [44], or BERT, is another breakthrough model based on the Transformer. It combines the benefits of ELMo's bidirectional training with GPT's Transformer

architecture, producing stellar results. BERT, like ELMo and the OpenAI Transformers, is a language model that pretrains on raw text and can be fine-tuned for many different tasks. But unlike a standard language model, which predicts a word given the sequence of preceding words, BERT masks input words at random and trains to predict the masked words. The architecture of this "masked language model" is merely a stack of Transformer encoders. At the time of its publication in late 2018, it achieved state-of-the-art performance on 11 NLP tasks.

Several variants of BERT exist that are worth mentioning. RoBERTa is a model that uses the same architecture as BERT but with minor adjustments that improve performance. For example, RoBERTa trains on more data for a longer period of time. It also dynamically adjusts the masking scheme and trains on longer sequences. T5[172] and BART[115] are models that adapt BERT to be better suited for text generation. Both of them add a Transformer decoder to the BERT encoder, decoding the original input one word at a time. They also have a more elaborate masking mechanism that includes blanks and masks of longer spans of text.

2.10 Transfer Learning

Transfer learning [161, 210] is the strategy of improving performance on a task with limited training data by pretraining the model on some related data-rich task, then fine-tuning it on the main downstream task. [257] That way, the pretrained model can fine-tune more effectively on scarce data because it is initialized having already learned how to extract salient features from similar data. It had long since been proven to be a powerful strategy on computer vision tasks. A CNN classifier could be pretrained on a large dataset like ImageNet, say, then fine-tuned on a small dataset of radiology images. [95]

Transfer learning has more recently shown immense usefulness in NLP.[151, 180, 181] Language models in particular are often used for pretraining. Their importance can be attributed to two reasons: one, they train to develop an intuitive understanding of semantics and grammar, and two, they are unsupervised models that train on raw text data - a resource that is in enormous supply. Models can pretrain on gigantic corpora like the entirety of English Wikipedia. Once pretrained, network architectures for language models like ELMo and BERT can be easily adapted for classification tasks like spam detection, sequence tagging tasks like named entity recognition, generative tasks like abstractive summarization, and many other NLP methods.[154, 168] The recent success of many pretrained models over existing benchmarks cements transfer learning's status as an indispensible tool in contemporary NLP.

A recent work [163] studies transfer learning in biomedical NLP by introducing evaluation of BERT [44] and ELMo[168] on ten benchmarking datasets. Results suggest that pre-trained BERT models on PubMed abstracts and clinical notes could significantly improve model performance on the selected NLP tasks.

3 CLASSIFICATION AND PREDICTION

Text classification and prediction tasks in EHRs are critical to quickly processing thousands of large texts for clinical decision support, research, and process optimization. In this section, we cover main subtasks include general medical text classification, segmentation, word sense disambiguation, entailment; and some special subtasks for EHRs: EHR de-identification, medical coding and outcome prediction for research purposes.

3.1 Medical Text Classification

There has been a large amount of work in text classification for medical texts and clinical notes. Some existing works rely on traditional and classical methods including Support Vector Machines (SVMs) and K-Nearest Neighbor (KNN). Marafino et al. developed an SVM classifier which was able to identify a range of diagnoses and procedures in intensive care unit (ICU). [134] In the work by Khachidze et al., SVMs and KNNs were applied to classify the documents describing instrumental diagnostics records.[93] After de-identification and simple processing like tokenization and lemmatization, they performed feature selection and applied the classifiers.

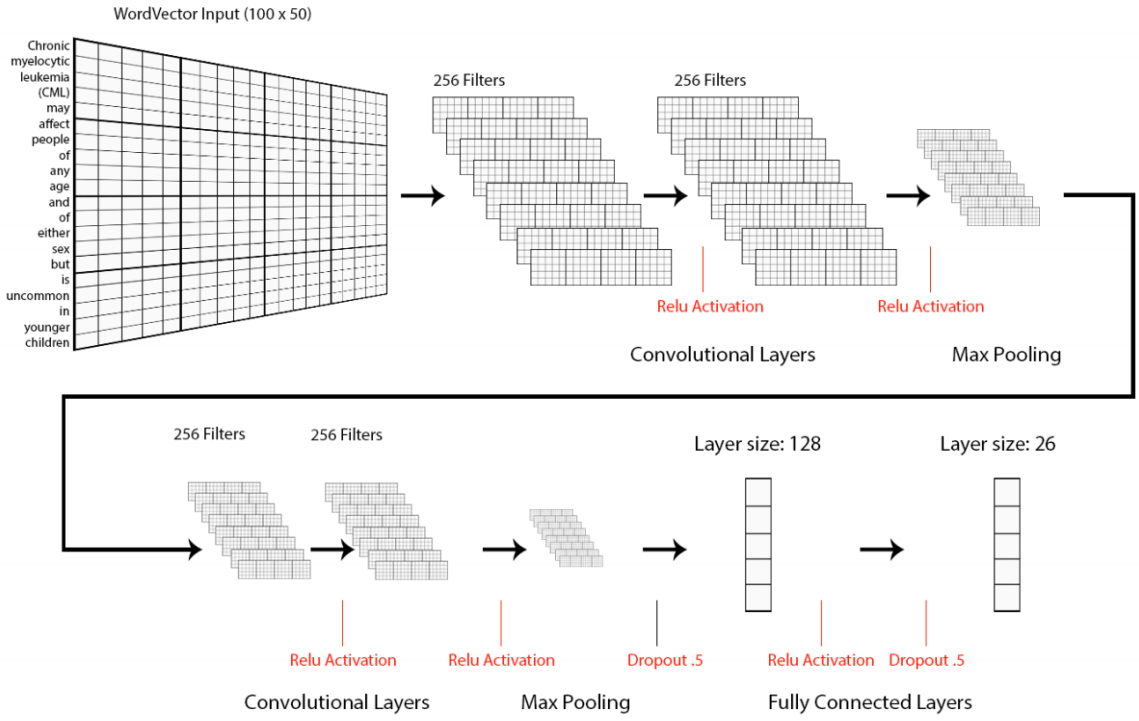

Fig. 1. Medical text classification using CNN. Adapted from the original paper. [77]

In recent years, some deep models like CNNs have attracted attention and achieved very competitive results in classification tasks. One of the very early attempts is by Hughes et al., who applied a CNN to classify clinical text at a sentence level. [77] The model structure is illustrated in Figure 1, where there are four convolutional layers after the sentence embedding input. Finally, a fully-connected layer is applied to predict the sentence labels. They compared their method to a variety of traditional machine learning methods and various sentence embeddings, including logistic regression, Doc2vec embeddings [108] and bag-of-word features. Their results show that the CNN-based classifier outperformed other classifiers in accuracy by about 15%. In other words, deep models including word embeddings and CNN models have been proven to outperform TF-IDF (Term Frequency - Inverse Document Frequency) and topic modeling features by a large margin. [77, 235] In a similar vein, Yao et al. introduced an approach to combine rule-based features and knowledge-guided deep learning techniques for disease classification using clinical texts. [254] These aforementioned approaches are supervised methods. When labeled data is not enough for supervised learning, weak supervision with pre-trained word embeddings is developed for clinical text classification.[235] Applying pre-trained models like BERT and BioBERT for classification is also an alternative. [136]

Besides the previously mentioned generalized medical text classification task, some work specifically focus on various applications in the medical domain:

Chief Complaint Extraction Models by Valmianski et al. extract the chief complaint described in the EHR. [219] This is essential for the development of systems that quickly triage patients without human intervention.

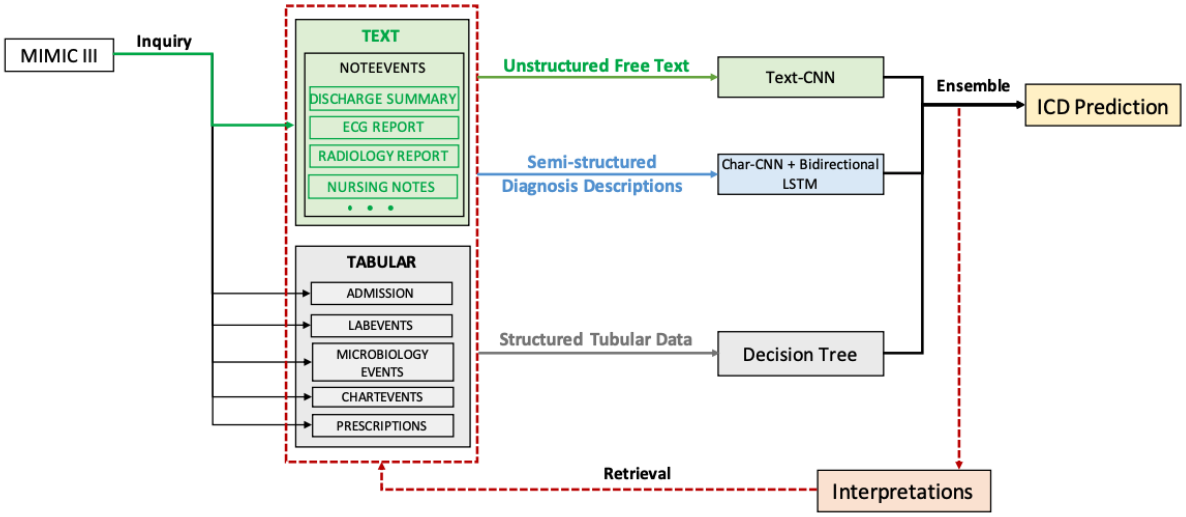

Fig. 2. A multimodal model architecture for ICD code prediction. [250]

They attempt this classification task with BERT-based and TF-IDF models. While TF-IDF models outperformed BERT-based models on this task, their robustness remains questionable.

Sepsis Detection Futoma et al. are able to use an RNN classifier to detect sepsis in EHRs. [58] Early detection is critical to improving patient outcomes, but is difficult to do because the symptoms overlap with other conditions. The data required to inform the prediction is already included in patient EHRs, so they frame the problem as a multivariate time series classification problem. They use a multitask Gaussian process (MGP) and feed into an RNN, and are able to make substantial improvements in performance compared to baselines and clinical benchmarks, and have significantly higher precision.

3.2 Medical Coding

The medical coding task attempts to map electronic health record text to International Classification of Diseases (ICD) codes. [187] Each of these codes represents a different diagnosis, and they are used in clinical treatment, medical billing, and statistics collections. ICD codes can help physicians quickly determine which diseases are involved and reach clinical decisions in a more timely manner, but are also tediously specific and detailed – diabetes alone has over two dozen different codes. Human coders struggle to manage the scale and complexity of the processes, often making mistakes and causing costly legal consequences in billing and beyond. There are many attempts to improve ICD coding using rule-based methods [53], or machine learning algorithms including Bayesian Ridge Regression [123], SVMs [100, 166], and so on.

Specifically, earlier deep learning-based methods have been focused on applying CNNs, LSTMs and LSTMs with attention by formulating this task as a document classification task. [10, 192, 250] A representative model with explainability is Convolutional Attention for Multi-Label classification (CAML). [153] It is a CNN-based model with attention: the model first aggregates information of the whole document with a CNN, then the attention mechanism is applied to select the most relevant segments from the document which trigger the ICD code prediction, providing possible explanations. Xu et al. [250] proposed a multimodal framework which considers unstructured texts, semi-structured texts and tabular data when predicting an ICD code. As shown in Figure 2 , a CNN is applied for modeling the unstructured texts; a deep learning model which contains a character level-CNN

and a Bi-LSTM is applied for semi-structured text; and finally a decision tree is applied for the tabular data. By assembling those models, the system is able to make ICD code predictions.

Focusing on features extracted from discharge summary notes in MIMIC-III, a widely used, albeit limited, EHR dataset, Huang et al. conduct a study on multi-label ICD-9 code classification.[75] The authors tested 1) a CNN-based classification model with word2vec document representations for each discharge summary, 2) a standard LSTM-based model, and 3) a GRU-based model on the sequential discharge summary text. The authors conducted two main classification tasks: predicting the top 10 ICD-9 codes, and predicting the top 50. First, on the top 10 codes, compared to several baseline methods, including Logistic Regression, Random Forest, and feed-forward networks, RNN-based methods showed greater performance in prediction. However, in the top 50 codes, Logistic Regression outperformed the deep methods proposed by the authors, leaving room for improvement in deep-learning methods for multi-label ICD-9 coding.

In a work using more recent methods, Singh et al.[198] implement a BERT model to predict ICD-9 codes from unstructured clinical text. They treat diagnosis and procedure code identification as a multi-label classification task, and use a BERT model trained on the MIMIC-III de-identified EHR database. As a result of using a bidirectional pre-trained language model, the model can learn the medical data context with less computation and preprocessing than other methods. BERT fine-tuned with MIMIC-III significantly out-performs existing literature, performing better as the classifier handles more diagnosis and procedure codes.

ICD coding can suffer from a severe data imbalance between the most and least popular codes. So Vu et al.[228] take an approach by moving to an attention-based model to adapt to the varied lengths of text fragments that caused challenges with prior convolutional neural network approaches. The label attention model consists of four layers: a pre-trained embedding layer, a bidirectional LSTM, an attention layer for label-specific weight vectors, and label-specific binary classifiers. All are tested on MIMIC-III, and when compared against state-of-the-art baselines in logistic regression, SVM, and CNNs, the label attention model substantially outperforms each and the hierarchical joint learning architecture succeeds in improving performance for rare codes.

At a lower level than documents, physicians may also need to automate coding at the encounter level. Shing et al. [195] propose encounter-level document attention networks (ELDAN), which are made of three parts: a document-level encoder that turns sparse document features into dense document features, a document-level attention layer, and an encounter-level encoder. They treat the problem as multiple one-vs-all binary classification problems, and are able to outperform the baseline for 17 of the 20 most frequent codes; all of this is achieved without needing to train on document-level annotations.

Schmaltz et al.[188] propose a zero-shot sequence labeling method that generalizes to the document-level multi-label setting, extending prior work where the labels were only binary. They use exemplar auditing, leveraging CNN filters to offer insight into the training set for a nearest neighbor to a relevant local feature in a test prediction so as to better evaluate the prediction. In order to evaluate the model, they focus on the ICD-9 medical coding task for the MIMIC-III dataset, and are able to produce competitive results with existing approaches.

3.3 Medical outcome prediction

Prediction of medical outcomes (e.g. death, progression to heart failure) is often difficult and aided by using information available within providers notes, imaging reports, etc. We highlight several recent results for neural NLP methods. Doctor AI[30] is a modelthat predicts all the new diagnoses and medications for a patient's subsequent visit. The researchers use an RNN because its sequential architecture lends itself to the temporal nature of a patient's healthcare trajectory. The input at each time-step is a raw representation of a single patient visit, incorporating relevant disease, medication, and procedure codes. The hidden state serves as a representation of the patient's medical history at that point in time. It outperforms existing baselines on differential diagnosis tasks. In a similar temporally-based prediction task, Suresh et al. propose LSTM and CNN-based models for

forward-facing predictions of ICU intervention tasks, including ventilation, using vasopressors, and using fluid boluses.[207] The data is split into 6 hour chunks, where patient data is recorded, as well as the status of the interventions being taken. After a 6 hour gap period, a 4 hour prediction period is allocated to test the model and predict which interventions were taken during this period. The authors report high AUC (area under curve) scores compared to previous models. DeepCare[169] is an end-to-end neural network that addresses the episodic nature and irregular temporality of electronic medical records. It reads medical records, stores illness history, infers current illness states and predicts future medical outcomes. Each care episode is represented with a vector; the relationship between them is modeled with a modified LSTM that accounts for irregularities in time, admission methods, diagnoses, and interventions. DeepCare aggregates the patient's medical history with a technique called multi-scale temporal pooling, then predicts the probability of some medical outcome. It demonstrates improved performance on predicting future diabetic and mental health outcomes. Moreover, Lyu uses a seq2seq model as both an autoencoder and forecaster, and shows that the integrated attention mechanism could improve clinical predictions. [129] Recently, considering the low-resource nature of clinical risk predictions for a given disease, Zhang et al. propose MetaPred, a transfer-learning framework to assess clinical risk for low-resource clinical disease data.[266] Using similar disease prediction datasets, MetaPred is trained to learn a classifier for the target disease. MetaPred is then fine-tuned on the target disease training data. As expected, in this low-resource setting, this framework outperforms methods using just the target disease data.

In a task closely related to outcome prediction, Hsu et al. address the predictive power of the variety of unstructured notes within EHRs.[74] Because some EHR unstructured notes include heavy copying from the structured fields, the authors investigate how valuable unstructured notes are for prediction tasks, and evaluate which parts of notes are most valuable. Using MIMIC-III, the authors consider readmission prediction and in-hospital mortality prediction to determine the value of unstructured notes. The authors find that . Also, they find that selected sentences from clinical notes can outperform using the entire set of notes for downstream prediction tasks.

Sometimes machine learning models are not interpretable and cannot correspond to or add to existing medical knowledge, making the model just an unknown black box. However, some following efforts shows that some models are interpretable, for example some models prioritize more recent visits than ones further back in time.

Two recent studies [21, 239] use high-performance generalized additive models with pairwise models to apply to pneumonia risk prediction and 30-day hospital readmission study. The pneunomnia case discovers surprising patterns in the data while achieving state-of-the-art result as the previous models. Some other efforts are also done to ensure neither accuracy nor interpretability is compromised. Choi et al. propose Reverse Time Attention Model (RETRAIN) [31], which makes it more interpretable by using a two-layer attention model, attending EHR data in a reverse time order, so that recent clinical visits receive a higher attention, and significant clinical variables within those visits. Trained on 14 million visits over an 8 year period, it achieves satisfying performance. Another approach to increase interpretability on models is suggested by Che et al. [23] The work introduces a knowledge-distillation approach known as interpretable mimic learning. The model uses gradient boosting trees to learn interpretable models from some existing deep learning models. Evaluated on an ICU dataset for acute lung injury, the model achieves similar or better performance than the deep learning ones.

Focusing on heart failure, kidney failure, and stroke, Liu et al. create models using unstructured and structured EHR data to predict these diseases.[125] In particular, they experiment with LSTMs, CNNs, and a combined CNN-LSTM hierarchical model to predict the onset of diseases prior in order to diagnosis time by using historical data occurring before a short gap period. Using a large de-identified hospital EHR dataset, the authors create their own domain-specific word embeddings to use in the deep models. A Bidirectional LSTM model performed the best across all disease prediction tasks. Additionally, in an comparative study, the authors find that integration of unstructured clinical notes data is critical for improving model performance.

Surprisingly, non-EHR data such as financial data also helps prediction of disease evolution. Sousa proposes a model predicting diabetes accessing high-risk diabetics, especially about acute complications (such as amputations and debridements), using financial data taken from healthcare providers. The model makes use of self-attentive recurrent neural networks with an input layer trained with word2vec skip-gram connected to a bidirectional LSTM. The model achieves the prediction of complications from 60 to 240 days with area under ROC curve from 0.81 to 0.94. [203]

3.4 Segmentation

Segmentation is the task of finding boundaries between sections of a text. The application of automatic segmentation in clinical documents is very important, as most EHRs of any length are either explicitly or implicitly segmented. In EHRs, topic segmentation aims to segment a document into topically similar parts, it is also known as text segmentation or discourse segmentation in various scenarios.

There are many attempts using traditional machine learning methods to perform segmentation. A work by Apostolova et al. proposed a model for segmentation of biomedical documents. [5] Using a hand-crafted training dataset of segmented clinical notes, they used a Support Vector Machines (SVM) to classify each sentence into a section, taking formatting and contextual features into account. Impressively for its time, their model achieved 90% accuracy. Similarly, an effective system was proposed which utilizes the sequential aspect of the sections of a clinical note, as it's more likely to go from certain section types to others. [119] They used a Hidden Markov Model on a labeled dataset of over 9,000 clinical notes to infer the state - the section label - of a span of text. Instead of a Hidden Markov Model, a 2012 work by Tepper et al. [213] favored classifying likelier EHR section sequences by using beam search and the Maximum Entropy approach. [17]

Rule-based systems have also been used for the task of clinical segmentation. Edinger et al.[47] develops rules from unstructured textual discharge summaries, MD notes, radiology reports, and nursing notes in the MIMIC-II database. Within each of these four types of EHRs, the researchers identified the most common section headings and structured the documents in XML format for future use. Their evaluation showed that this rule-based segmentation method produced strong results.

However, there are limited works attempting to utilize deep models for segmenting medical texts. Word embeddings are investigated to improve the segmentation for medical textbook chapters. [7, 92] The attention mechanism has been applied for a neural segmentation model.[7] The segmentation task is then formulated as a classification task by predicting whether each sentence in a text is the beginning of a section. For each sentence, they use CNNs to create sentence embeddings for both the sentence in question and its context sentences on each side. Next, the middle sentence attends to these sentence embeddings to create context vectors. Finally, these representations are merged and used for binary classification. The model improves over contemporary baselines in WinDiff scores [186] on three benchmark datasets - one of which is a clinical text dataset containing 227 chapters from a medical textbook. [133]

3.5 Word Sense Disambiguation

Word Sense Disambiguation (WSD) aims to assign the correct meaning to an ambiguous word given its context. In the medical domain, EHRs often contains many ambiguous terms that are involved with specific domain knowledge. For example, the word "ice" may refer to frozen water, methamphetamine (an addictive substance), or caspase-1 (a type of enzyme). [236] The WSD task has been approached with supervised learning, semi-supervised learning and knowledge-driven methods. [55, 124, 244] These approaches show that massive high-quality annotated training data is essential to achieve desirable WSD system performance. This is especially true for medical WSD, where only experts with substantial background knowledge can annotate. In addition

Note 1
...72 year-old male with history of DM2 (Diabetes Mellitus Type 2), myocardial infarction requiring CABG(coronary artery bypass graft), asthma, MR, and germ cell tumor with metastases to left upper lobe...

Note 2
...She also underwent an echocardiogram which showed left ventricular systolic function which was normal. She had mild MR and mild TR. She had some early diastolic dysfunction as well as biatrial enlargement...

Fig. 3. An example for abbreviation disambiguation, adapted from the original work. [117]

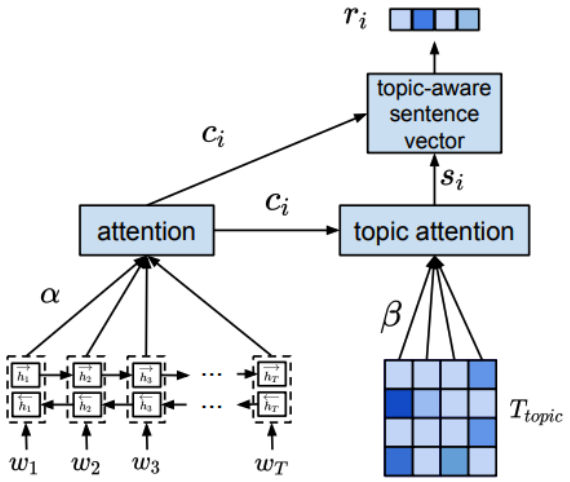

Fig. 4. The topic-attention model for abbreviation disambiguation. [117]

to improving understanding of the contents of an EHR, WSD can also help in downstream tasks like machine translation, information extraction, and question answering. [22, 173, 271]

While some works applied machine learning models to tackle WSD, for example, Bayesian algorithm, SVM, Naive Bayes, and decision trees) they need a certain amount of labeled data. [20, 113, 248]

Recently, neural methods are investigated for WSD. The deepBioWSD model [167] is a representative work. First, they utilized the Unified Medical Language System (UMLS) sense embeddings; then these embeddings are applied for initializing a single bidirectional long short-term memory network (Bi-LSTM), which is then trained to do sense prediction for any ambiguous term. Other works investigated similar structures, i.e. multi-layer LSTMs [18], Bi-LSTM with self-attention [261] and so on.

3.5.1 Abbreviation Disambiguation. A special case of WSD, the task of medical term abbreviation disambiguation, is essential for people to improve their understanding of medical records. As with WSD, this task can also assist other downstream NLP tasks like sentence classification, named entity recognition, and relation extraction. [83] In clinical notes, it is quite common for physicians, nurses or doctors to apply medical term abbreviations to represent drug names, disease names and other words. Depending on the medical specialty and contents of the EHR, these abbreviations can have a wide range of possible choices.[88, 244] For example, there exist at least 5 possible word senses for the term MR, including magnetic resonance, mitral regurgitation, mental retardation, medical record and general English Mister (Mr.).

Some efforts have been made for abbreviation disambiguation on clinical notes. Traditional methods like decision trees are applied for acronym disambiguation in Spanish EHRs. [179] Xu and Stetson [249] were the first to apply clustering techniques in building word sense inventories of abbreviations in clinical text. Later on, efforts were made to utilize word embeddings for abbreviation disambiguation. A work [245] examined three methods for word embeddings from unlabeled clinical corpus: an existing method called Surrounding based embedding feature (SBE), and two newly developed methods: Left-Right surrounding based embedding feature and MAX surrounding based embedding feature.

In a recent work, Joopudi and Dandala [88] propose a convolutional neural network (CNN) model that encodes representations of clinical notes and predicts abbreviation sense as a classification task. Their model is shown to be robust across different abbreviation datasets. A neural topic-attention model is applied to learn improved contextualized sentence representations for medical term abbreviation disambiguation. [117] In this method, shown in Figure 4, a Latent Dirichlet Allocation (LDA) model was leveraged to learn topic embeddings, then contextualized word embeddings (ELMo) are applied to conduct a topic-aware sentence vector for classification. Besides, Adams et. al [2] propose the Latent Meaning Cells (LMC) model for clinical acronym expansion, focusing on utilizing meta-data and lexical context for contextualized representation.

3.6 De-identification

The task of removing the patient sensitive information and preserving clinical meaning in EHR data is referred as de-identification. De-identification of PHI (protected healthcare information) in EHR clinical notes is a critical step to protect patient information before sharing or publishing datasets for secondary research purpose. The Health Insurance Portability and Accountability Act (HIPAA) includes 18 different types of protected health information (PHI) [1], i.e., names, locations, phone numbers and so on. Several shared tasks have been organized to promote de-identification of clinical text in the NLP field. [205, 206, 218] Given the large amount of EHR data needed, manual de-identification is not feasible given the amount of time it requires. Rule-based de-identification studies mainly depend on dictionary pattern matching and regular expressions[206]. However, these methods usually require complete algorithms and cannot handle out-of-expectation cases such as typos and infrequent abbreviations.

These days in NLP, it is common to utilize automatic de-identification approaches that apply named entity recognition (NER) methods. Some other automatic de-identification tools use hybrid approaches to combine rules-based method and machine learning NER methods[206] to reach good performance. But these method still have difficulties when applied to data in another language, or when used in different clinical domains. Recently, there have been many attempts to improve automatic de-identification using deep learning.

The early de-identification systems for patient records based on artificial neural networks proposed to apply structures such as LSTMs and LSTM-CRF models. [43, 253] The system consists mainly of four components: an embedding layer, a label prediction bidirectional LSTM layer, a CRF (conditional random field) layer and a label-sequence optimization layer. These systems usually outperformed a CRF-based approach by a large margin. Moreover, the CRF-based approach needs manual feature engineering from domain exports.

Besides, some studies found that the embedding layer has a large impact on the model performance. [246] found that an integration RNN model with medical knowledge from Unified Medical Language System outperforms a baseline RNN model. Another study[211] shows that the BERT embedding has better overall performance than any?(which model) other embedding method.

The problem of de-identification is notably more difficult in non-English medical texts, in which datasets are even more limited. Many studies apply the mature approaches for English text to EHRs in other languages. Trienes et al.[214] evaluates existing de-identification methods for their performance across two languages

[1] https://www.hhs.gov/hipaa/index.html

and three clinical domains. They compare performances of a rule-based system for Dutch psychiatric EHRs, a feature-based CRF, and a deep neural network (BiLSTM-CRF) for transferability and generalizability from English to Dutch. They found that the deep neural network performed best, interestingly beating the Dutch psychiatric de-identification model, but still performs worse when applying a pre-trained model in new domains. García-Pablos et al.[160] tested a pre-trained multilingual BERT model with several Spanish clinical texts. They compared the BERT-based sequence labelling model with a baseline sensitive data classifier, the spaCy Spanish NER model, and CRFs. They were able to show that the simple BERT-based model without domain-specific fine-tuning is able to out-perform all other methods and is additionally robust to training-data scarcity. Kajiyama st al.[91] applies rule-based, CRF, and LSTM-based methods on three Japanese EHR datasets. They observe the LSTM-based method has the best performance and robustness between different sources.

4 CLINICAL EMBEDDINGS

This section explores various instances of representation learning in the EHR domain. Such representations have been used to model the semantics of biomedical text, the health trajectory of individual patients, and many other important tasks. We also introduce how recent pre-trained language models could promote better representation learning in EHR and biomedical text.

4.1 Concept Embeddings

Medical concepts contain various types, for example, genes, proteins and diseases. To learn the semantics of these medical concepts could be helpful for machine learning tasks in the medical domain.

The Cui2vec[11] framework learns from a huge corpus of insurance claims, clinical notes, and journal articles. This corpus contains 80 million documents and over 100,000 medical concepts. As a preprocessing step, the researchers map each medical concept word or phrase to its corresponding "concept unique identifier," or CUI. The co-occurrence statistics they glean from these mappings are used to construct concept embeddings with GloVe and word2vec.

Predictive modeling in healthcare is hindered by data insufficiency and a lack of interpretability. The GRaph-based Attention Model (GRAM)[33] addresses both of these issues by supplementing limited EHRs with relevant hierarchical information from medical ontologies. It embeds concepts with a weighted combination of their ancestors in the ontology via an attention mechanism. That way, each concept embedding is informed by relevant background knowledge. The concept embeddings for a visit are joined to form a visit representation, which is then used for prediction. GRAM is shown to yield impressive predictive accuracy when trained on small amounts of data. Figure 5 shows some example concept embeddings trained using GRAM model.

Many concepts, including lab tests, diagnoses and drug administrations, are temporal in nature. These "heterogeneous temporal events" may have a high degree of correlation between them. For example, the event of some diagnosis being made is strongly correlated to the results of certain lab tests. In addition to understanding the relationships between events, the model must also represent their temporal nature. Some medical events happen only once, whereas others will occur periodically according to some visiting rate. Ignoring the dynamic nature of embedded units may lead to issues with the semantics of learned representations. Liu et al.[126] developed a model for obtaining the joint representation of heterogeneous temporal events. It is trained on the overarching classification task of clinical endpoint prediction, which predicts whether some medical event like a disease or symptom will happen in the future. Their model's main contribution is a modified LSTM cell based on Neil et al.'s Phased LSTM.[158] The Phased LSTM alters the traditional LSTM by adding a time gate that accounts for inputs with irregular sampling patterns. Liu et al.'s model goes a step further by adding an event gate, which is capable of modeling correlations between thousands of event types.

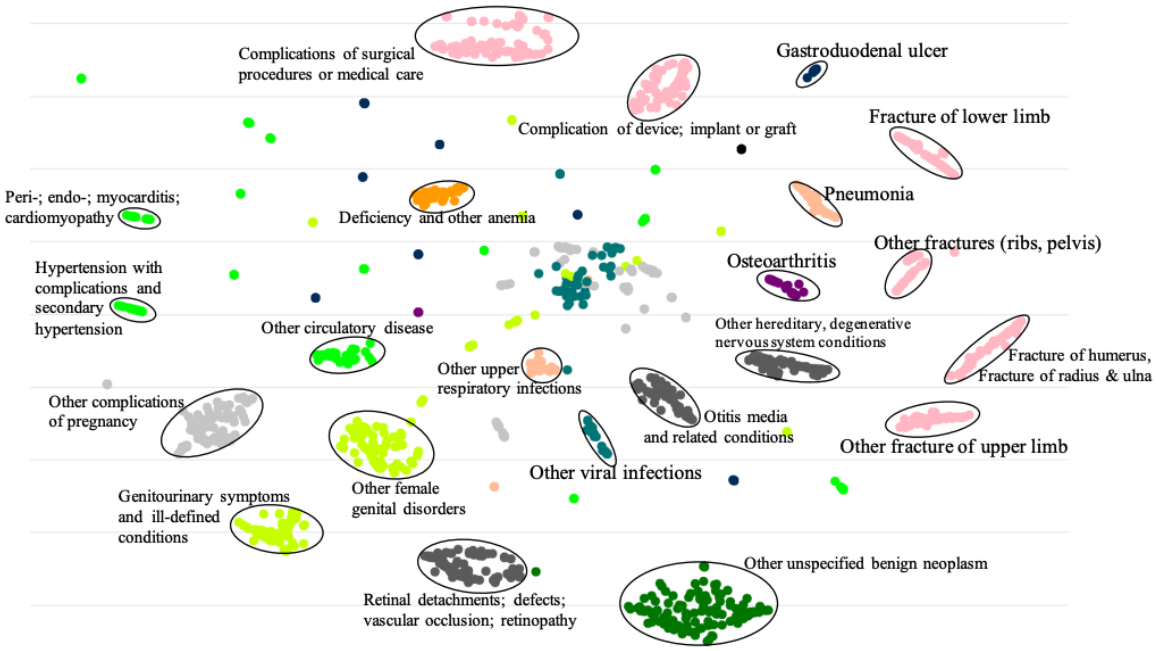

Fig. 5. t-SNE Scatter plot of the medical concepts trained using GRAM. Adapted from the original paper. [33]

Zhu et al. [272] also preserves temporal properties by compressing each patient visit into a fixed-length vector with medical embeddings based on surrounding medical context. These event embedding vectors are stacked together to produce a dense embedding matrix for each patient. Pairs of these matrices are passed through convolutional filters and mapped to feature maps. These feature maps are then pooled into intermediate vectors to build the embedding patient representations. A symmetrical similarity matrix is constructed using the distance between the feature vectors. Their proposed framework achieves strong performance on similarity measuring among patients compared to baseline approaches.

4.2 Visit Embeddings

Med2vec[32] generates vector representations of patient visits and medical concepts like procedures, diseases, and medications. It represents each patient visit as a vector of medical codes corresponding to the concepts that occurred in the visit. Like the Word2vec skip-gram model, Med2vec inputs the current visit, embeds it, and predicts the likelihood of previous and future visits. Doing so utilizes the sequential order of a patient's visits and the co-occurrence between concepts. Once trained, the model outputs concept embeddings by simply inputting one-hot vectors. Its parameters can be interpreted via qualitative inspection by identifying the inputs for which they are highest, then finding patterns of disease group or visit type within these inputs. Med2vec embeddings outperform previous models at tasks like predicting concepts in future visits and calculating the patient's current severity status.

Most models represent each patient visit as a flattened collection of concepts like diagnoses and treatments. Doing so, however, ignores valuable hierarchical information on the multilevel relationship between the concepts in an EHR. For example, the diagnosis of a fever can lead to treatments like acetaminophen and IV fluid, which can

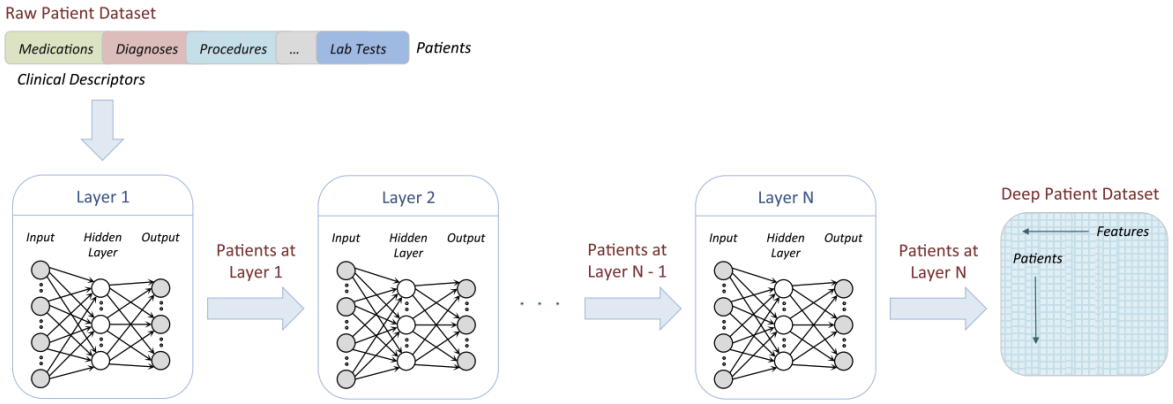

Fig. 6. The diagram of the Deep Patient framework, adapted from original paper[144]: an unsupervised deep learning method which is trained on raw dataset, and is able to learn patient representation in the last neural network layer.

in turn produce side effects that need treatments of their own. MiME[35] models the inherent hierarchical structure of EHRs, leveraging it to create embeddings at the concept, visit, and patient level while jointly performing auxiliary prediction tasks. These multilevel embeddings predicted heart failure and sequential diseases with greater accuracy and/or less data than existing methods.

The neural clinical decision support system by Wei et al.[238] creates visit representations in a simple yet clever way. First, they extract diagnostic ICD codes from the MIMIC-III database. With these as labels, they train a CNN to predict the patients' codes from the raw EHR text. The final dense layer of the network is taken to be a visit representation. This representation is successfully applied to the information retrieval task of recommending relevant literature for individual patients. Here, visit representations from MIMIC aid in achieving strong performance on a task for which little training data exists.

The deep neural network by Escudié et al.[50] learns low-dimensional representations of patient visits when ICD codes are removed to predict the presence or absence of such codes. A CNN applied to the text features and a multi-layer perceptron (MLP) used on the structured data are trained together. The last hidden layers of each subnetwork are concatenated to produce an embedding for each stay. This embedding is able to conserve semantic medical representation of the initial data and improves on the prediction performance of a random forest classifier.

4.3 Patient Embeddings

Mehrabi et al.[140] represent patient notes with a matrix. The rows of the matrix represent medical codes, and the columns represent years. Patient embeddings are created from this matrix with an unsupervised network called a deep Boltzmann machine. This model is highly restricted, however, because the only information it utilizes is the 70 most frequent ICD-9 codes. In addition, all temporal information is chunked in units of one year, which may be too long in the medical context.

Deep Patient [144] is a 2016 patient representation framework published in Nature. First, it extracts raw features like ICD-9 codes, medications, lab tests, and concepts from preprocessed EHRs. Each patient is represented with either single vector or by a sequence of vectors determined by temporal windows. It then embeds these raw vectors with a stacked denoising autoencoders. Figure 6 shows the framework diagram. These patient embeddings

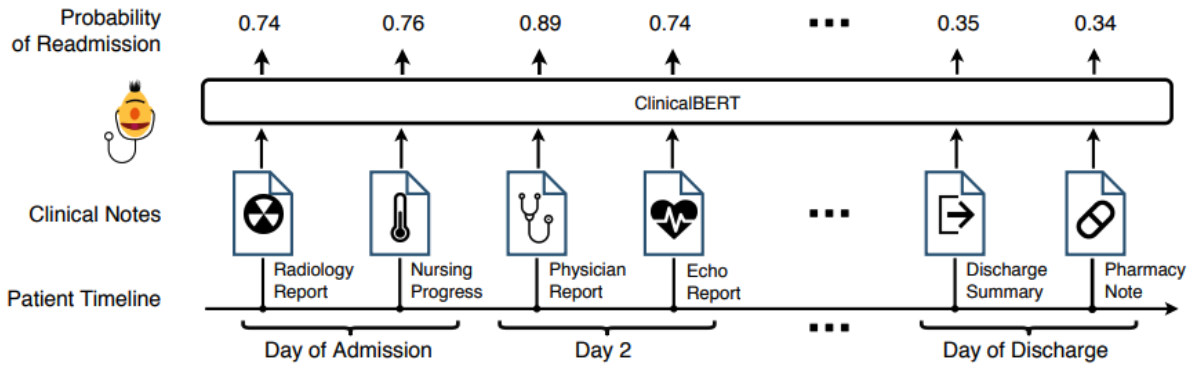

Fig. 7. ClinicalBert illustration[76]: notes were added to electronic health record during a patient's admission to update the patient's risk of being readmitted within a 30-day window.

are applied to the task of clinical disease prediction, and shown to improve over existing baselines. They were especially effective at predicting diabetes, schizophrenia, and various cancers.

Dligach and Miller [45] build upon methods in Deep Patient to learn patient representations using only text variables. First, the neural network model takes a set of CUIs as input and produces a vectorial representation of the patient. The final network layer is composed of sigmoid units that are used to jointly predict all possible billing codes associated with the patient. The learned dense patient representations were successful in outperforming sparse patient representations on average and for most diseases.

A more recent model, Patient2vec[262], represents each patient visit as a sequence of its ICD-9 codes, medications, and lab tests. It learns an embedding for each of these codes by using Word2vec to predict the codes that are likely to co-occur in a visit. Then, it represents a patient's whole history in a single embedding with an RNN and attention mechanism. Patient2vec predicted future hospitalizations with higher statistical power than previous patient embedding models.

Sushil et al.[208] learn unsupervised patient representations directly from clinical text. They attempt this with two neural approaches: a stacked denoising autoencoder and a paragraph vector architecture called doc2vec.[109] Doc2vec, like word2vec, embeds each word based on the other words that appear in a context window. However, it adds a document- or paragraph-level vector that it embeds alongside the other words. These neural networks produce patient representations that outperform other embedding models at tasks like predicting mortality and primary diagnostic category. In addition, they are useful at demonstrating similarity between patients.

TAPER[40] uses text and medical codes to produce a unified representation from a patient's visit data that can be used for downstream tasks. The medical code embedding is learned by a skip-gram model using transformer networks, while a pre-trained BERT[44] model produces the medical text embeddings. These two embeddings are concatenated for the final patient representation. TAPER demonstrated better results in prediction on tasks of readmission, mortality, and length of stay compared to other methods such as Med2vec and Patient2vec.

4.4 EHR Embeddings

BERT[44] harnessed the power of Transformers to generate better word embeddings than ever before. However, its embeddings generalize poorly to text from specific domains like biomedicine. For this reason, several papers subject the pre-trained BERT to a subsequent round of pre-training - this time, on corpora of domain-specific text. Gu et al.[63] demonstrated that such domain-specific pretraining from scratch can yield stronger results

than mixed-domain pretraining, including pretraining from general-domain language models. BioBERT[111] adapts BERT for biomedical text, training on biomedical research papers from the Pubmed corpus. SciBERT[12] trains BERT on 1.14 million biomedical and computer science articles from the Semantic Scholar corpus. And, most relevantly to EHRs, ClinicalBERT[76] trains on clinical notes from the MIMIC-III corpus. Figure 7 shows the illustration of ClinicalBERT. In this particular example, the notes are added to EHR, in order to update the patient's risk of being readmitted within a 30-day window.

EhrBERT[116] also trains on clinical notes, but it is not generally available because its training dataset is not public. These models all outperform BERT on tasks in their domains.

Jin et al.[85] generate embeddings from scientific text by training ELMo on 10 million PubMed abstracts. They call this model BioELMo and benchmark it against BERT, ELMo, and BioBERT. Surprisingly, embeddings from BioELMo outperform those from BioBERT on a number of tasks like named entity recognition.

MT-Clinical BERT[154] goes a step further than these models. In addition to learning embeddings of clinical text, it also performs multitask learning on eight information extraction tasks including entity extraction and personal health indicator (PHI) identification. The BERT embeddings are shared as inputs to these prediction tasks. This multitask system is competitive with task-specific information extraction models, as it shares information in an efficient manner.

BEHRT [120] is a deep neural transduction model that learns about patients' past diseases and the relationships that exist between them. It uses BERT's masked language models pretraining approach but relies on the four key embeddings of diseases, age, segment, and position. This produces a final embedding that preserves timing of events along with data concerning disease sequences and delivery of care. Evaluation of the model demonstrated BEHRT's superior predictive power in disease trajectory compared to other powerful approaches such as RETAIN[31].

MS-BERT [38] is a transformer model trained on real clinical data, rather than the MIMIC corpus. The model is trained on over 70,000 Multiple Sclerosis (MS) consult notes and publicly available on line. [2] Before training, the notes are de-identified. Then the model is tested on a classification task to predict Expanded Disability Status Scale (EDSS), which is usually inside unstructured notes. The model surpasses other models that applied word2vec on this task by a large margin.

CheXbert[199] applies BERT to the task of labeling free-text radiology reports. Existing machine learning methods in this task either employ feature engineering or manual annotations from experts. While high-quality, the annotations are in short supply and expensive to create. CheXbert overcomes this limitation by learning to label radiology reports with both annotations and existing rule-based systems. It first trains to predict the outputs of a rule-based labeler, then fine-tunes on an augmented set of expert annotations. It sets a new state-of-the-art for report labeling on a large dataset of chest x-rays.

5 INFORMATION EXTRACTION

Information extraction (IE) is the task of automatically extracting information from unstructured natural language text. It encompasses several subtasks, including named entity recognition (NER), event extraction, and relation extraction (RE). In this section, we present an overview of various IE tasks and methods as applied to EHRs.

5.1 Named Entity Recognition

NER is the task of determining whether tokens or spans in a text correspond to specific types of "named entity," of interest, such as people, medication, and diseases. For example, in the sentence "Paris is the capital of France," "Paris" would be classified as a city and "France" would be classified as a country. But if the sentence were "Paris is a media personality and scion of the Hilton family," "Paris" would be classified as a person.

[2] https://huggingface.co/NLP4H/ms_bert

Given the importance of contextual information for this task, models that use bidirectional context tend to produce better results at this task. Gilgic et al.[60] showed that medical NER performance could be improved by first pretraining word embeddings on unannotated EHRs with word2vec[142]. As far as embedding models go, however, BERT[44] is more sophisticated than word2vec. Because of this, the creators of BioBERT[111] trained the normal pre-trained BERT on biomedical text, then fine-tuned it for a number of NER tasks. BioBERTs embeddings proved very effective at recognizing entities such as diseases, species, proteins, and adverse drug reactions. Peng et al.[163] similarly adapted BERT for biomedical text; their version outperformed BioBERT on NER tasks like recognizing diseases, chemicals, and disorders.

While effective, these BERT models fail to address a central issue in medical NER: the transferability between specialties. Clinicians in different medical specialties use vastly different vocabularies, which poses a huge challenge in training an effective one-size-fits-all medical NER model. This problem is exacerbated by the scarcity of publicly available data, especially for certain specialties. Wang et al. approach this issue by developing a double transfer learning framework for cross-specialty NER.[237] This system transfers both feature representations and parameters, which enables resource-poor specialties to utilize knowledge gleaned from specialties with sufficient numbers of annotated EHRs.

5.2 Event Extraction

The goal of event extraction is to detect different types of events of interest and their properties. HYPE, for example, is a system that predicts whether sentences in an EHR contain a hypoglycemic event. [86] Jagannatha and Yu use bi-directional LSTMs and GRUs to extract acute myocardial infarction (AMI) cases from free-text in EHRs, providing references to other salient information such as medical background, family history, and diagnosis date. [80] Their results demonstrate that deep learning methods can yield improved performance over more traditional machine learning models like that of Zheng et al.[269], which uses conditional random fields for the AMI detection task. Event extraction approaches can generally be used for other types of information, such as symptoms, and are not necessarily distinct from NER approaches. Du et al. [46] proposed novel models to extract symptoms mentioned in clinical conversations along with their status. Two neural models are introduced including a hierarchical span-attribute tagging model and a sequence-to-sequence model that decodes the symptoms and their status.

5.3 Concept Extraction

As discussed previously, EHRs include several *concepts* that provide critical information on a patient's medical conditions and treatment trajectory. Extraction of medical concepts is an active area of research. A 2018 paper by Tao et al. embeds each word in an EHR note, then uses the embeddings to predict whether they denote a medical concept.[212] Each embedding is a concatenation of two separate representations: one derived from ELMo, an unsupervised bidirectional language model, and the other from a publicly available medical ontology. The latter enables the model's embeddings to draw from a larger knowledge base that contains domain knowledge. The model performs prediction with a conditional random field, which takes neighboring tokens into account when classifying each word.

A 2020 system [103] extracts diagnoses and organ abnormalities from doctor-patient conversations. These conversations are often too long to extract information from directly, so the researchers first filter utterances by how "noteworthy" they are. From these utterances, they use a BERT-based model to recognize the important medical concepts.

5.4 Medication Information Extraction

EHRs are a trove of vitally important information pertaining to medications. Amazon recognized the importance of the medication IE task, investing heavily in its Amazon Comprehend Medical system. This system is able to extract information about a patient's condition, medication, frequency, dosage, and strength. Its model encodes an EHR with two separate LSTMs, then extracts concepts with a tag decoder.[66] Another model, RNNG, extracts medication names and information from EHRs by using an RNN to encode sentences as constituency trees.[114] This grammar-aware approach proved effective at identifying relations between medical entities.

Like Amazon Comprehend Medical, Mahajan et al. also use NER in their model for medication dosage extraction. Theirs, however, is the first to automatically compute daily dosage. [132] They utilize a BERT-based NER system to extract various forms of medication information from the text: names, dosages, units, frequencies, etc. The model then uses this information to calculate the daily dosage of each medication.

A framework by Selvaraj and Konam extracts dosage and frequency information from doctor-patient conversations, rather than EHRs.[191] They framed this problem as a Question Answering (QA) task, in which the question was the conversation and medication name, and the answer was the medication's dosage and frequency. To accomplish this, they used a pointer-generator seq2seq model with separate decoders for dosage and frequency.

5.5 Entity Linking

Entity linking is a slightly more refined task than NER. Unlike NER, which links each mention to a broader class, entity linking links each mention to a specific entity. In the earlier example of "Paris is the capital of France," while an NER model would place the mention "Paris" in the general class of "city," an entity linking model would be more specific. It would link it to a particular entity: the city of Paris, France. As before, one word can refer to several potential entities, so context matters. In this case, "Paris" could be the city Paris, France or the person Paris Hilton or even the city of Paris, Illinois. These ambiguities also exist in biomedical text. In the sentence fragment, "symptoms of cold include cough and runny nose," the mention "cold" must be linked to one of many possible disease entities.

Traditional clinical entity linking models detect mentions and identify a list of candidate concepts for them. MEDTYPE[220] goes a step further, incorporating an entity disambiguation step to filter out unlikely candidate concepts. This step predicts the semantic type of an identified mention based on its context. For example, to continue with the above example, "cold" has the semantic type Disease/Syndrome. Correctly identifying this helps filter out wrong entities like "Cold Temperature." To train MEDTYPE, its authors introduce WikiMed and PubMedDS, two large datasets of medical entity mentions. As of May 2020, it is the best-performing medical entity linking model.

The entity linking task has important implications for relation extraction. The same two words can have completely different relationships with each other depending on their semantics. For example, the words "model" and "train" could be related in reference to a model of a train or the training a machine learning model. Thus, the entity predicted by an entity linking model affects the relation predicted by a relation extraction model. This can lead to a chain of cascading errors.

Bansal et al. avoids this problem by developing a single joint model for entity linking and relation extraction.[8] Their model SNERL considers all possible graphs of entities and relations, and predicts the most likely one. This method performs favorably compared to the existing state-of-the-art pipeline approach. Moreover, it does so without requiring any mention-level entity annotations.

5.6 Relation Extraction

The task of relation extraction extracts the entities in a text that share a given relation. This is critical in our context because an NLP system must grasp the relationships between various medical entities in order to fully

understand a patient's health. A number of these important relations exist in EHRs. The creators of BioBERT, for example, fine-tuned it on the GAD and EU-ADR datasets to extract gene-disease interactions. [111] They also extracted chemical-protein interactions with the ChemProt dataset.

Later in 2019, Peng et al. also pre-trained BERT on PubMed articles and fine-tuned it for named entity recognition and relation extraction. [163] Their models, however, included only the paper abstracts in the training data; some also trained on MIMIC-III clinical notes. Their relation extraction models achieved higher F1 scores than BioBERT on the DDI, ChemProt, and i2b2 2010 datasets. These models extracted interactions between drugs, chemicals, proteins, medical problems, tests, and treatments.

Datta and Roberts created a model to extract spatial relations from radiology reports. [41] In their paper, they present the following sentence as an example: "There are areas of airspace opacity within the left lung base which may represent atelactasis or infiltrate." From this sentence, "within" is a spacial indicator linking the trajectory, "airspace opacity," with the landmark, "left lung base." To extract this, the model trains on the Rad-SpRL dataset, using Bidirectional LSTMs to generate syntax-aware word representations.

Medical relation extraction, like many other NLP tasks, suffers from a lack of training data. One of several shortages is on relations between proteins. For example, only 4% of protein-protein interaction (PPI) have annotations of function in the IntAct database. PPI-BioBERT[48] leverages information in EHRs to augment existing PPI datasets with an ensemble of BioBERT relation-extractors.

Similarly, more unlabeled information on drug-drug interactions (DDI) exists in EHRs than in existing medical resources. DDI is an extremely important interaction to detect, as taking certain drugs simultaneously can produce adverse effects. To take advantage of this, RE models like the attentive LSTM from Sahu et al.[184] automatically extract drug-drug interactions directly from free-text medical records.

One recent paper shows the success of multi-task learning on these RE tasks. The multi-task learning framework from Yadav et al.[252] jointly models PPI and DDI. While it has not yet been peer-reviewed, the authors claim it "significantly" improves over existing medical RE baselines.

6 GENERATION, SUMMARIZATION AND SIMPLIFICATION

In this chapter, we introduce recent breakthroughs of clinical text generation, summarization and simplification.

6.1 Generation

Clinical text generation is the Natural Language Generation (NLG) task of creating novel clinical text from preexisting data.[78] This generation task is particularly important in the medical domain due to accessibility and confidentiality issues surrounding EHR data. Even de-identified data still risks re-identification via the residual data. Therefore, in generating high-quality clinical data, researchers can test NLP techniques in the medical domain without the overhead privacy measures and agreements necessary with de-identified datasets. Medical report generation has the ultimate goal of reducing the administrative burden and complexity of material that physicians must write. Research has focused more on generating paragraphs of clinical text, but generated medical reports may also in some cases improve clinical outcome, identifying characteristics in images that physicians may have bypassed.

A frequent application of medical report generation is captioning medical images. Writing medical reports is a task that is tedious for experienced radiologists, and challenging for newer radiologists who do not yet possess the requisite skills. The model must create a paragraph description of the image, both accurately identifying all abnormalities and generating complex sentences, which are longer and more complicated than the usual natural image captions. Li et al. propose an Auxiliary Signal-Guided Knowledge Encoder-Decoder (ASGK) that attempts to mimic radiologists' working patterns.[118] They take advantage of two types of auxiliary signals to do so, the internal fusion features and external medical linguistic information. The medical graph and natural language

decoders are pretrained with external auxiliary signals to memorize and phrase medical knowledge, and train with internal signals to support the graph encoding which integrates prior medical knowledge and visual and linguistic information. The AGSK is able to outperform other state-of-the-art methods in report generation and tag classification on the CX-CHR dataset and a new COVID-19 CT report dataset as measured by focal loss.

In a different direction, medical report generation can also be used to support other research. Electronic health records are notoriously limited due to sensitive patient information–common de-identification techniques are not always robust against re-identification attacks, so one solution to create datasets is simply to generate synthetic EHRs.

Choi et al. propose an approach to synthetic EHR generation via a medical Generative Adversarial Network (medGAN).[34] medGAN generates high-dimensional multi-label discrete variables to represent variables found within an EHR, such as medications, diagnoses, and procedures. Via a combination of an autoencoder and GAN, the model is trained to learn the distribution of these discrete high-dimensional EHR variables. The autoencoder is learned from a source EHR input, and the decoder from the autoencoder is used to convert continuous outputs from the generator into a discrete form. The discriminator then makes judgements on source data and generated data passed from the generator and decoder. Besides several outliers, a trained medical professional found records from medGAN and the source data relatively indistinguishable.

Baowaly et al. improve upon the existing medical generative adversarial network (medGAN) method, proposing medWGAN and medBGAN.[9] The medWGAN model uses the WGAN-GP model as the generative network, which employs gradient penalties to overcome sample quality challenges, and medBGAN uses the BGAN model, where the generator is trained to create samples on the decision boundary of the discriminator. medWGAN yields the highest precision and medBGAN yields the highest recall for both datasets, and both significantly surpass medGAN alone. As this field progresses, the body of EHR datasets can be artifically generated to further improve the quality and quantity of training data so that other spaces in EHR NLP research can also move forward.

In earlier work on generation, Lee introduces an artificial clinical text generation method via a feed-forward encoder and RNN decoder.[112] The source EHR dataset includes over 5 million deidentified health records, and this study focused on generated artificial chief complaints, a freetext part of EHRs. In their model, the encoder takes into account many patient variables, like age, disposition, and diagnosis, and attempts to decode this into a textual chief complaint. They find that their generated complaints are largely epidemiologically valid, and preserve the relationships between the diagnoses and chief complaints.

Recognizing the need for more work on artificially generated clinical notes data, Melamund and Shivade introduce a generation task, including a dataset derived from MIMIC-III, privacy measures, and utility benchmarks for generated text.[141] They use an LSTM language model to generate a synthetic dataset mimicking the style and content of the original dataset. The privacy measure they introduce ensures that individual information is not deducable from aggregate statistical information from a dataset. The measure ensures that the information contained in the aggregate dataset is similar to the aggregate dataset missing one record.

In a more domain-specific setting, Hoogi et al. focus on generating artificial mammography reports via an LSTM architecture trained on real mammography reports.[72] These types of models input an image like an X-ray and generate a description of what it shows. Their work introduces an interesting qualitative analysis in which they ask a radiologist to distinguish between real and generated reports; the radiologist classified real reports correctly 86% of the time, and fake ones as real 75% of the time, suggesting that the fake reports are of high quality. They show that augmenting real data with generated data significantly improves performance in a downstream benign/malignant breast tumor classification task.

In October 2020, a Stanford team[147] also published a work on radiology report generation. They address the problem that generated reports tend to leave out critical information or hallucinate knowledge on the image's contents by introducing two new metrics for this task. The first of these evaluates the report's completeness by measuring the extent to which the radiology domain entities cover their ground-truth references. And the

second metric further evaluates the factual consistency by using an NLI model to measure the generated reports' logical entailment with their references. They optimize these two metrics in a reinforcement learning framework, leading to a significant improvement over report generation baselines.

Recently, Amin-Nejad, Ive, and Velupillai proposed a generation technique using an encoder-decoder Transformer model in a seq2seq framework.[4] Using the MIMIC-III database, they model text generation with information of the patient and the ICU stay as inputs, and the textual discharge summary as outputs. They compare the generation capacity of the vanilla Transformer model to that of GPT-2. With more data, the Transformer model produces more realistic text, but GPT-2 performs better in a data-scarce setting, an important result given the restrictions present around real EHR data. They use generated artificial data in an augmentation scheme and test its utility by training a model for downstream tasks, including readmission prediction and phenotype classification. They find that this augmentation boosts performance on downstream tasks.

6.2 Summarization

Nurses, doctors and researchers deal with massive EHRs daily. In this case, text summarization could reduce their workloads to condense the document into a brief, readable summary.[146] As a fundamental NLP task, there have been many works in summarization on general domain, such as news articles, [49, 52, 190], scientific papers [1, 255] and dialogues. [259] Recent attempts have included the EHR domain; we group these works into extractive summarization and abstactive summarization.

6.2.1 Extractive Summarization. Portet et al.[170] proposed one of the first attempts at extractive summarization of EHRs. Their model was designed for neonatal intensive care data, which consists of both free text and discrete information (e.g. equipment settings and drug administration). Handling these diverse types of data in one model is a challenging task, and human evaluators found model summaries to be unhelpful.

For this reason, most subsequent attempts handled only textual data. One 2019 graph-based model ranked sentences by the important biomedical concepts they share. [150] To circumvent the shortage of labeled training data, Liu et al. performed extractive summarization by utilizing intrinsic correlation between EHRs within a disease group to generate pseudo-labels.[127]

Query-based summarization models produce a summary of sentences relevant to some input query. EHRs are a strong target for these systems because they would enable a physician to quickly search for relevant medical information. McInerney et al.[138] design a query-focused extractive summarization model that provides sentences significant to a potential diagnosis. Because no large corpus of EHRs with extractive summaries exists, they use a distant supervision framework that extracts ICD diagnosis codes from future visits. They train a transformer-based neural network to select the summary sentences from an EHR and use them to predict future diagnoses.

Another model that produces extractive summaries from queries directly compares the query sentence with each candidate sentence from the EHR[149]. An embedding model like BERT embeds these sentences. A strong benefit of this system is that it is capable of performing multiple-document summarization, rather than being limited to a single document.

6.2.2 Abstractive Summarization. More training data exists for abstractive summarization. For this task, models can use the "impressions" section of a clinical note as the labeled abstractive summary of the "findings" section, which constitutes the main text. Four recent models train and evaluate seq2seq models to summarize radiology reports, using an additional context vector that provides the decoder with relevant background information. With the exception of Sotudeh et al., each model has a modified pointer-generator network architecture. Two of them link entities in the input to rich domain-specific information from medical ontologies like UMLS. [59, 131]

Background: radiographic examination of the chest ... Findings: continuous rhythm monitoring device again seen projecting over the left heart. persistent low lung volumes with unchanged cardiomegaly. again seen is a diffuse reticular pattern with interstitial prominence demonstrated represent underlying emphysematous changes with superimposed increasing moderate pulmonary edema. small bilateral pleural effusions. persistent bibasilar opacities left greater than right which may represent infection versus atelectasis.
Human Summary: increased moderate pulmonary edema with small bilateral pleural effusions. left greater than right basilar opacities which may represent infection versus atelectasis.
Baseline Model Summary: no significant interval change.
Zhang Model Summary: increasing moderate pulmonary edema. small bilateral pleural effusions. persistent bibasilar opacities left greater than right which may represent infection versus atelectasis.

Fig. 8. Example result of Zhang's RL-based abstractive summarizer.[268] It is capable of achieving near-human performance.

The other two, both led by Yuhao Zhang from Stanford, summarize the radiology report's "Background" section. [267] The more recent model adds two new features: a fact-checking mechanism and a reinforcement learning (RL) objective. [268] It computes a factual correctness score between the model summary and the CheXpert labeler [79], which extracts fact variables from a source radiology report. The model aims to optimize an RL objective that balances the model summary's factual accuracy, linguistic likelihood, and overlap with the target summary.

A separate area of research is concerned with summarization of questions and their answers. Abacha and Demner-Fushman[13] introduce a corpus of summarized consumer health questions and use it to train an effective pointer-generator network for abstractive summarization. Similarly, Savery et al.[185] develop a dataset of common consumer health questions, their answers, and summaries of their answers. They demonstrate its usefulness by successfully employing a number of state-of-the-art extractive and abstractive summarization models on it. The Biosquash system takes a different approach: rather than generating answers or summarizing existing ones, their model summarizes multiple documents relevant to a question. [193]

6.3 Simplification

EHRs are inaccessible to most readers by nature. Doctors write them with complex medical jargon that is unfamiliar to the layperson. This makes them a prime target for NLP models that perform text simplification. That way, people would be able to easily read complex medical documents; terms like "peripheral oedema" would be replaced or tagged with "ankle swelling." We focus on two NLP methods for simplification: Lexical Substitution and Neural Text Simplification.

6.3.1 Lexical Substitution. Lexical Substitution (LS) is the task of replacing complex words in a text with simpler paraphrases. It is comprised of two subtasks: Complex Word Identification (CWI) and selecting the best paraphrase for the word given its context. Researchers at UPenn achieved state-of-the-art performance in their 2018 LS system. [104] For CWI, they use a Support Vector Machine (SVM) to classify each word as complex or non-complex. The model takes into account a word's number of characters, syllables, definitions, synonyms, and occurrences. It also

considers the average of these values for the rest of the words in the sentence, presuming that complex words are likely to be in a sentence with other complex words. They then consider candidate substitutions for each word predicted complex from Simple PPDB, a database of 4.5 million pairings of complex words with simpler paraphrases. [162] They select the paraphrase whose skip-gram word embedding most closely resembles that of the complex word. [142]

SimpleScience is a unified framework for CWI and paraphrase selection for text in the scientific domain.[97] Because models that perform well on certain types of text often generalize poorly on scientific text, SimpleScience applies well to EHRs. Its approach is to generate word embeddings with Word2Vec on both a scientific and a general corpus. It also calculates the complexity of words in each corpus based on their frequency and length, under the rationale that more complex words tend to be longer, occur frequently in the scientific corpus, and occur rarely in the general corpus. To perform lexical substitution, the model selects the word whose embedding in the general corpus has highest cosine similarity with that of a more complex word in the scientific corpus. It evaluates on a hand-crafted scientific LS dataset called SimpleSciGold, finding greatly improved performance over existing models.

6.3.2 Neural Text Simplification. Rather than replacing individual words, Text Simplification models are seq2seq frameworks that rewrite entire passages. As one might imagine, these are more expressive but difficult to train. Most are trained on Newsela, which paraphrases news articles with 5 separate levels of complexity, or pairings of Wikipedia articles with their Simple Wikipedia counterparts. Many measure success with the 2016 SARI benchmark, which stands for **S**ystem output **A**gainst **R**eferences and against the **I**nput sentence.[251] A high SARI means that the n-grams in the system output closely resemble the reference simplifications, but not the input text.

Neural Clinical Paraphrase Generation (NCPG) is a recent model that casts clinical paraphrasing as a monolingual neural machine translation problem.[67] Using a character-level, attention-based bidirectional RNN in an encoder-decoder framework paradigm to NMT efforts, NCPG outperforms a baseline word-level RNN encoder-decoder model. Models were evaluated on a constructed dataset combining Paraphrase Database 2.0 and a medical thesaurus, Unified Medical Language System, to build a clinically oriented parallel paraphrase corpus. In addition to evaluating performance metrics of the model, the authors show that the character-based NCPG model is superior to word-level based methods as it tackles the out-of-vocabulary problem directly.

Zhang and Lapata, 2017, used a Reinforcement Learning (RL) framework for sentence-level text simplification. Their model, DRESS, uses an encoder-decoder RNN with attention to produce generated simplifications. These simplifications are scored by a reward function that encourages simplicity, relevance, and fluency. It trains the model parameters using REINFORCE, a classic RL policy gradients algorithm. [242]

Vu et al.'s model improved over DRESS on the TS task.[227] Its major development was using a Neural Semantic Encoder (NSE) rather than a standard LSTM encoder. [155] The NSE is a memory-augmented RNN architecture that utilizes the entire source sequence instead of just the RNN's previous state. Because of this, it produces better sentence-level embeddings on long sequences.

Different readers require different levels of simplification. For example, a physician in a different specialty or a hospital administrator wouldn't require as simple paraphrasing as an average patient. Recent work in Text Simplification addresses this by developing models that permit the user to choose their output's level of complexity. But complexity can be measured in different ways - someone who is dyslexic, for example, wouldn't need the same kinds of simplification as someone for whom English is not their native language. The ACCESS model, thus, lets the user select a number of attributes such as lexical complexity, length, amount of paraphrasing, and syntactic complexity. [135] It uses a Transformer architecture and achieves state-of-the-art performance on the SARI benchmark.

The above methods are supervised approaches. There are a few unsupervised methods proposed for text simplification. Weng et al. perform unsupervised text simplification specifically for clinical notes - an important development that circumvents the shortage of labeled simplifications in this domain. [239] They get skip-gram embeddings from two clinical corpora: MIMIC-III, which is full of medical jargon, and MedlinePlus, which is oriented for the layperson. A Bilingual Dictionary Induction model is used to align these embeddings of technical and simpler terms and initialize a denoising autoencoder. This autoencoder inputs a physician-written sentence, generates a simpler translation with a language model, and uses back-translation to reconstruct the original sentence.

There is also a growing demand of clinical notes translation from medical jargons into layperson understandable terms to enhance patient's understanding in their own conditions. Existing approaches used dictionary-based replacement, which are limited to expert annotations and not scalable. Weng explores the clinical word and sentence translation in a completely unsupervised manner. [239] They discovered that a combination of representation learning, bilingual dictionary induction and machine translation yields the best precision.

7 OTHER TOPICS

7.1 Question Answering

Question answering (QA) is the task of interpreting natural language questions and retrieving appropriately paired answers. General QA systems have had recent success with pre-trained language models, but biomedical QA faces domain-specific challenges. The primary barrier for biomedical QA in EHRs is domain-specific vocabulary. Models trained on general domain corpora have difficulty understanding biomedical questions because of the hyper-specific technical vocabulary often used in clinical settings.

7.1.1 Transfer Learning Methods. A recent work [191] applied QA technique to extract the Medication Regimen (MR, dosage and frequency for medications) discussed in a medical conversation. They formulate the MR task as a QA task and generate questions using the template: "What is the <dosage/frequency> for <Medication Name>? ". They used a pointer-generator network [190] with a co-attention encoder.

Another approach to biomedical QA is to take advantage of large pre-trained models. BioBERT[111], a pre-trained biomedical language model trained on PubMed articles, has been successfully adapted for the QA task. The pre-trained language model needs to be fine-tuned on biomedical QA datasets. However, biomedical QA datasets are often very small (limited to just a few thousand samples) and limited in scope, and creating new datasets is frequently cost-prohibitive. So they first fine-tune BioBERT on large-scale general domain extractive QA datasets, and then fine-tune on the biomedical BioASQ dataset.[215] Using this transfer learning framework, they are able to significantly outperform the basic BERT [44] and other state-of-the-art models in the QA task, as well as other biomedical NLP tasks, improving and overcoming challenges in both range of vocabulary and dataset size. Vilares et al.[223] recognize the same problem with limited datasets and focus on the task of multi-choice QA, which requires knowledge and reasoning in complex domains. They create their own dataset, HEAD-QA, from the Spanish government's annual specialized healthcare exams. They then translate the dataset to English to perform cross-lingual experiments and compare evaluations for an IR model on the Spanish HEAD-QA and cross-lingual models for English HEAD-QA. Being able to train with cross-lingual datasets or potentially automatically translating question-answer pairs could open the door to improvements in multilingual QA so that answers are not limited to only the query's native language.

7.1.2 Problem Transformation. It is also possible to transform the QA task as other tasks including question similarity, information retrieval, and recognizing question entailment.

Question similarity and question entailment have been promising paths to solving the biomedical QA task. Medical questions are asked much more frequently online than can be answered, but will often take similar

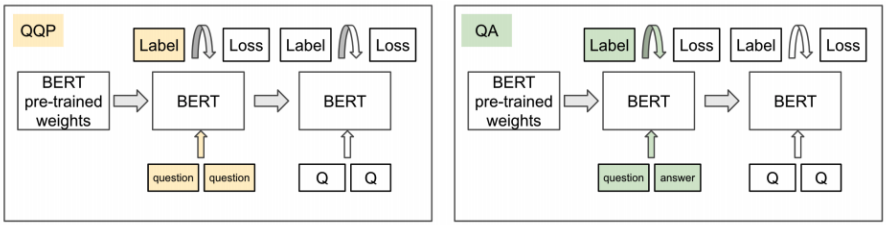

Fig. 9. The double finetune method using a pre-trained BERT to an intermediate task to medical question-similarity task for two different intermediate tasks: Quora question-question pairs (left) and medical question-answer pairs (right).[137]

forms; the goal of question entailment is to map new questions to similar answered questions. An in-domain semi-supervised approach was proposed and tested on 3,000 medical question pairs. [137] They pre-train BERT on the HealthTap dataset [3] and double fine-tune, first on intermediate tasks like Quora question similarity and medical answer completion and then on medical question pairs, illustrated in Figure 9. The original BERT with fine-tuned weights was compared against SciBERT, ClinicalBERT, and BioBERT, each also additionally fine-tuned on the final medical question pair task. BERT outperformed both SciBERT and ClinicalBERT, and was beaten by only BioBERT.

The same task can also be solved by combining IR models with recognizing question entailment (RQE) methods[14] . RQE is a task similar to natural language inference; it interprets sentences and attempts to extract meaning. However, RQE tries to create relevant relaxations of contextual and semantic constraints, such that specific questions can be related to more general and already-answered questions. A few works [14] attempt two approaches to RQE: 1) a neural network, and 2), a logistic regression classifier. The neural network performed best with the general domain NLI datasets, but logistic regression resulted in higher accuracy for the domain-specific datasets, specifically consumer-health questions which would be more applicable for general medical QA use.

Other attempts including applying paraphrasing are proposed to improve QA systems in EHRs. [201, 202] A representative work [201] collected 10,578 unique questions via crowdsourcing to train a model consisting of a variational autoencoder and LSTMs using the question paraphrasing method.

7.2 Phenotyping

Computational phenotyping is the process of extracting clinically relevant characteristics from patient data. These characteristics include physical traits, physiology, and behavior. Phenotyping is used in several areas of medical research, such as categorizing patients by diagnosis for further analysis and identifying new phenotypes.[260] Recent techniques in computational phenotyping have replaced traditional rule-based phenotyping algorithms with NLP models. However, some of these approaches require large amounts of labeled data. Therefore, most recently, research has turned to unsupervised learning for phenotyping; this technique can also give rise to novel phenotypes.

Zhang et al. propose an unsupervised deep learning model to identify phenotypes in EHRs. [263] They make use of the Human Phenotype Ontology, and assume the semantic latent representation of EHRs is a combination of the same representations for textual phenotypic descriptions. They use an autoencoder to first learn the semantic vector representations, and then the contributions of each phenotype representation to an overall EHR representation. They also use a classifier to ensure the learned representations are different enough from each

[3] https://github.com/durakkerem/Medical-Question-Answer-Datasets

other. Their approach achieves competitive results, and is much faster than previous phenotype identification models.

Another recent unsupervised approach, Granite, uses a tensor factorization method with limited human supervision, improving on classic dimensionality reduction techniques. [69] Granite is a robust Poisson Nonnegative tensor factorization model (NNTF) that encourages diverse and sparse latent factors. It introduces angular penalty and an L2 regularization term, reduces overlap between factors, and also introduces simplex projection on factors which results in better sparsity control. Empirical work shows Granite yields phenotypes with more distinct elements, and is better than previous tensor factoring methods at capturing rare phenotypes.

Chiu et al. introduce bulk learning to the infectious disease domain, which uses a small dataset to simultaneously train and evaluate a large amount of phenotypes. [28] This method uses diagnostic codes as surrogate labels and trains an intermediate model based on feature abstractions that capture common clinical concepts among multiple clinical conditions. Each disease can then be labeled by multiple clinical concepts. The training stage consists of three parts, firstly, base classifiers are trained to predict labels of the set of the infectious diseases. Next, predictors are aggregated through a meta-classifier, and returns a feature abstraction that describes the extent of effect that a base model has on the prediction on a disease. In the final stage, a small subset of disease cases are collected to produce an annotation set. In effect, bulk learning serves to separate disease batches while using less data annotations. It is an example of multiple classifier hierarchical learning.

7.3 Knowledge Graphs

Knowledge bases (KBs) are datasets that store large amounts of information about structured or unstructured data. KBs whose entities are interlinked by some semantic rule or rules are known as knowledge graphs. Because words are related via various kinds of relationships, knowledge graphs are indispensable resources in NLP. When properly tapped into, they can provide vitally important insights to computational models. For example, WordNet[217] is a popular knowledge graph from Princeton whose entities - words - are linked to one another by various semantic relations. The most commonly utilized relation between words in WordNet is synonymy. Each group of words that share a common definition are grouped together into a synset.

7.3.1 Types in Medical Domain. A number of knowledge graphs pertain specifically to biomedical text. Among these, several contain triples of biomedical entity pairs and their relations. These knowledge graphs are crucial sources of training data for biomedical relation extraction task. One such knowledge graph is the Comparative Toxicogenomics Database [107], which provides information on chemical–gene, chemical–protein, chemical–disease and gene–disease interactions. In addition, the Human Phenotype Ontology (HPO) knowledge base connects diseases with pheonotypic abnormalities. [99]

Another important kind of interaction is disease-symptom. This relationship can be the basis for a knowledge graph in which a disease shares an edge with a symptom if it is a potential cause. Such a knowledge graph would be hugely useful because it would provide doctors and patients alike with a simple way to link symptoms with diseases, rather than having to sift through medical textbooks or dozens of websites. It would also be enormously popular: according to product manager Prem Ramaswami, one in 20 Google searches is for health-related information.[174] Figure 10 provides an example of part of a disease-symptom knowledge graph.

7.3.2 Building from EHRs. INTERNIST-I[143] and its successor, Quick Medical Reference[122], were two early knowledge graphs of medical entities. Developed in the 20th century, they were manually curated graphs that required hundreds of thousands of clinician-hours to build. Shwe et al. estimated that constructing Quick Medical Reference took no fewer than 15 clinician-years - a tremendous expenditure of time and money [196]. Moreover, these systems were brittle against new medical findings, as they could only be updated with more manual effort.

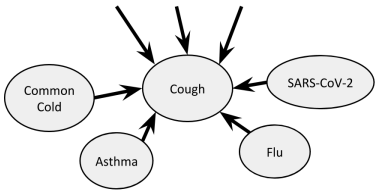

Fig. 10. A "cough" node of a knowledge graph that links diseases with symptoms.

Yet this trend of spending huge amounts of resources to develop medical knowledge graphs by hand continued until the 2015 release of the Google Health Knowledge Graph[174].

Finally, in 2017, Rotmensch et al.[178] proved the feasibility of automatically learn a knowledge graph from EHRs. For their model's training data, they extracted ICD-9 codes of diseases and UMLS codes of symptoms from a dataset of over 270,000 medical records. They used this data to fit three probabilistic models for disease prediction. Lastly, they built the knowledge graph by linking diseases with the symptoms for which the models indicated the highest probability.

Another study [27] presents a new methodology for analyzing the performance of EHR-based knowledge graphs, evaluating Rotmensch's system against the Google Health Knowledge Graph. It finds that automated EHR systems generally perform worse on diseases with more co-occurring diseases, more co-occurring symptoms, and fewer samples. It also finds that model performance varies widely on demographic information like age and gender.

While this task of using EHRs to automatically construct knowledge bases could have large implications, little research has been conducted on it. And to the extent of our knowledge, as of now no attempts have used neural networks for it.

7.4 Medical Dialogues

Recent natural language understanding (NLU) research on doctor-patient dialogues has large potential implications. The two primary focuses are automatic scribing and automatic health coaching.

Automatic scribing is extremely valuable because many physicians today spend hours dealing with administrative tasks like filling in information for electronic health records; in fact, one estimate has physicians spending two hours on EHRs for every hour spent with a patient. The simple solution is to hire a medical scribe–the medical scribe takes notes about the patient-physician encounter to reduce the physician's administrative burden and ensure that documentation in the electronic health record is accurate and up-to-date. However, scribes are a costly solution–roughly $49,000 onsite and $24,000 remote. [19] To solve this problem, NLP researchers are trying to automatically generate clinical notes from medical dialogue.

Automatic health coaching tries to reactively generate dialogue on top of transcription and interpretation. For simple questions that do not require complex or nuanced guidance, a health coach can be a cost-effective solution for many who cannot afford usual healthcare system costs.

7.4.1 Automatic Scribing. One model, AutoScribe, automatically parses the dialogue for entities like medications, symptoms, times, dates, referrals, and diagnoses; It then uses this information to generate a patient note. [94] While AutoScribe produced strong results, the researchers noted that it could be improved by extracting more entities and training on more dialogues. More recently, Krishna et al.[102] also tackled this task of generating summaries of medical dialogues. Their best-performing model extracts important utterances from the dialogue and clusters them into subsections; from this organizational schema, it effectively generates a structured SOAP (Subjective, Objective, Assessment, and Plan) note. To aid this task, Yim et al.[256] created a useful dataset and

annotation methodology. Their methodology associates each note sentence with a set of corresponding dialogue sentences. Sets with similar or related information can be grouped together further, forming a more sophisticated labeling scheme.

7.4.2 Conversational Coaches. Others models attempt the NLP task of developing conversational agents in the medical domain. Personal health coaches are useful but inaccessible for lower-income patients, so Gupta et al.[65] developed an automatic health coach that texts patients via SMS. This health coach communicates important medical information, sets concrete goals, and encourages the patient to adhere to them. Another model, from Campillos-Llanos et al.[128], learns to simulate a patient, not the medical professional. They develop a "virtual patient" dialogue system with which a physician can practice clinical interactions.

7.5 Multilingual

There is a recent interest in processing non-English medical texts, which often are less available than ones written in English. [159]

Roller et al. propose a sequential cross-lingual candidate search method for biomedical concept normalization. [177] The main component for the model is a neural machine translation network trained on UMLS for Spanish, French, Dutch and German. The proposed model performs similarly well compared with commercial translators (Google, Bing) on these four languages. Perez et al. compare the effectiveness of various approaches in automatic annotation of biomedical texts in Spanish. [165] The first is information retrieval and concept disambiguation. The second one is machine translation, which annotates documents in English and translates them back into Spanish. A hybrid approach combining the above two is also explored, and they find that the hybrid approach is the best out of the three.

BERT model for other languages rather than English is also investigated for EHR tasks. Vunikili et al. study BERT-based embeddings trained on general domain Spanish text for tumor morphology extraction in Spanish clinical reports. [229] The model achieves promising results on the NER task without any feature engineering or rule-based methods. Silvestri et al. study the multilingual Transformer-based model, Cross-lingual Language Model (XLM), [36] and evaluate in a cross-lingual ICD-10 classification on short medical notes. [197]

7.6 Application in Outbreak of Public Health Crisis

The COVID-19 pandemic, as a public health crisis, is largely impacting people's life in many perspectives. It is also an information crisis with the development of the Internet and other techniques. In NLP domain, researchers focus on processing pandemic-generated data and work on a various tasks including information retrieval, named entity recognition, knowledge discovery, etc. [26]

TREC-COVID [175, 226] is an information retrieval shared task to promote and support research related to the pandemic. Among the participants, MacAvaney et. al [130] introduce a zero-shot SciBERT-based ranking algorithm for COVID-related scientific literature. Bendersky et. al[16] present a weighted hierarchical rank fusion approach. The approach ensembles results from lexical and semantic retrieval systems, pre-trained and fine-tuned BERT rankers, and relevance feedback runs.

COVID-19 Open Research Dataset Challenge (CORD-19) corpus [231]

CORD-NER [234]

COVID-KG [232] is a knowledge discovery framework focusing on extracting multimedia knowledge elements from 25,534 peer-reviewed papers. In COVID-KG, nodes are entities/concepts and edges are relations and events among these entities. The edges are extracted from both images and texts. As a result, the knowledge graph contains 7,230 Diseases, 9,123 Chemicals and 50,864 genes, 1,725,518 chemical-gene links, 5,556,670 chemical-disease links, and 7,7844,574 gene-disease links.

8 DATASETS AND TOOLS

8.1 Datasets

8.1.1 General Datasets. **MIMIC-III**[4] A free, publicly accessible database of de-identified medical information on patient stays in the critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The table NOTEVENTS contains clinical notes from over 40,000 patients. Other tables have data on mortality, imaging reports, demographics, vital signs, lab tests, drugs, and procedures. Before MIMIC-III, there were two other iterations of MIMIC used by biomedical NLP researchers.

MIMIC-CXR[5] Like MIMIC-III, MIMIC-CXR contains de-identified clinical information from the Beth Israel Deaconess Medical Center. It has over 377,000 radiology images of chest X-rays. The creators of the dataset also used the ChexPert[79] tool to classify each image's corresponding free-text note into 14 different labels.

NUBes-PHI [121] A Spanish medical report corpus, containing about 7,000 real reports with annotated negation and uncertainty information.

Abbrev dataset[6] This dataset is a re-creation of an old dataset which contains the acronyms and long-forms from Medline abstracts. It is automatically re-created by identifying the acronyms long forms in the Medline abstract and replacing it with it's acronym. There are three subsets containing 100, 200 and 300 instances respectively. [204]

MEDLINE[7] A database of 26 million journal articles on biomedicine and health from 1950 to the present. It is compiled by the United States National Library of Medicine (NLM). MedlinePlus[8], a related service, describes medical terms in simple language.

PubMed[9] A corpus containing more than 30 million citations for biomedical and scientific literature. In addition to MEDLINE, these texts come from sources like online books, papers on other scientific topics, and biomedical articles that have not been processed by MEDLINE.

8.1.2 Task-Specific Datasets. **BioASQ**[10] An organization that designs challenges for biomedical NLP tasks. While BioASQ challenges focus primarily on question answering (QA) and semantic indexing, some use other tasks including multi-document summarization, information retrieval, and hierarchical text classification.

BIOSSES [11] [200] A benchmark dataset for biomedical sentence similarity estimation.

BLUE [12] Biomedical Language Understanding Evaluation (BLUE) is a collection of ten datasets for five biomedical NLP tasks. These tasks cover sentence similarity, named entity recognition, relation extraction, document classification, and natural language inference. BLUE serves as a useful benchmarking tool, as it centralizes the datasets that medical NLP systems evaluate on.

Clinical Abbreviation Sense Inventory[13] A dataset for medical term disambiguation. In the latest version, 440 of the most frequently used abbreviations and acronyms were selected from 352,267 dictated clinical notes.

CLINIQPARA[201] A dataset with paraphrases for clinical questions. Contains 10,578 unique questions across 946 semantically distinct paraphrase clusters. Initially collected for improving question answering for EHRs.

[4] https://mimic.physionet.org/
[5] https://physionet.org/content/mimic-cxr/2.0.0/
[6] https://nlp.cs.vcu.edu/data.html
[7] https://www.nlm.nih.gov/bsd/medline.html
[8] https://www.nlm.nih.gov/bsd/medline.html
[9] https://www.ncbi.nlm.nih.gov/guide/howto/obtain-full-text/
[10] http://bioasq.org/
[11] http://tabilab.cmpe.boun.edu.tr/BIOSSES/
[12] https://github.com/ncbi-nlp/BLUE_Benchmark
[13] https://conservancy.umn.edu/handle/11299/137703

i2b2[14] Informatics for Integrating Biology and the Bedside, or i2b2, is a non-profit that organizes datasets and competitions for clinical NLP. It has numerous datasets for specific tasks like deidentification, relation extraction, clinical trial cohort selection. These datasets and challenges are now run by Harvard's National NLP Clinical Challenges, or n2c2; however, most papers refer to them with the name i2b2.

MedICaT[15] A collection of more than 217,000 medical images, corresponding captions, and inline references, made for figure retrieval and figure-to-text alignment tasks. Unlike previous medical imaging datasets, subfigures and subcaptions are explicitly aligned, introducing the specific task of subcaption-subfigure alignment.

MedNLI[16] Designed for Natural Language Inference (NLI) in the clinical domain. The objective of NLI is to predict whether a hypothesis can be deemed true, false, or undetermined from a given premise. MedNLI contains 14,049 unique sentence pairs, annotated by 4 clinicians over the course of six weeks. To download it, one first needs to get access to MIMIC-III.

MedQuAD[17] Medical Question Answering Dataset, a collection of 47,457 medical question-answer pairs created from 12 NIH websites (e.g. cancer.gov, niddk.nih.gov, GARD, MedlinePlus Health Topics). There are 37 question types associated with diseases, drugs and other medical entities such as tests. [15]

VQA-RAD[18] A dataset of manually constructed question-answer pairs corresponding to radiology images. It was designed for future Visual Question Answering systems, which will automatically answer salient questions on X-rays. These models will hopefully be very useful clinical decision support tools for radiologists.

WikiMed and PubMedDS [220] Two large-scale datasets for entity linking. WikiMed contains over includes 650,000 mentions normalized to concepts in UMLS. PubMedDS is an annotated corpus with more than 5 million normalized mentions spanning across 3.5 million documents.

PathVQA [68] The first dataset for pathology visual question answering. It contains manually-checked 32,799 questions from 4,998 pathology images.

MedQA [84] The first multiple-choice OpenQA dataset for solving medical problems. The dataset is collected from professional medical board exams on three languages: English, simplified Chinese, and traditional Chinese. For the languages, there are 12,723, 34,251, and 14,123 questions respectively.

8.2 Tools and libraries

8.2.1 Machine Learning. **Pytorch** [19] Pytorch is an open source deep learning library developed by Facebook, primarily for use in Python. It is a leading platform in both industry and academia.

Scikit-learn[20] An open python library providing efficient data mining and data analysis tools. These includes methods for classification, regression, clustering, etc.

TensorFlow[21] Google's open-source framework for efficient computation, used primarily for machine learning. Tensorflow provides stable APIs for Python and C; it has also been adapted for use in a variety of other programming languages.

8.2.2 NLP. **AllenNLP**[22] An open-source NLP research library built on PyTorch. AllenNLP has a number of state-of-the-art models readily available, making it very easy for anyone to use deep learning on NLP tasks.

[14] http://www.i2b2.org/
[15] https://github.com/allenai/medicat
[16] https://jgc128.github.io/mednli/
[17] https://github.com/abachaa/MedQuAD
[18] https://osf.io/89kps/
[19] https://pytorch.org/
[20] https://scikit-learn.org/
[21] https://www.tensorflow.org/
[22] https://allennlp.org/

Fairseq[23] A Python toolkit for sequence modeling. It enables users to train custom models for text generation tasks like machine translation, summarization, and language modeling.

FastText[24] A sequence modeling toolkit which allows users to train custom models for translation, summarization, language modeling and other text generation tasks.

Gensim[25] A scalable, robust, efficient and hassle-free python library for unsupervised semantic modelling from plain text. It has a wide range of tools for topic modeling, document indexing, and similarity retrieval.

Natural Language Toolkit (NLTK)[26] A leading platform for building Python programs concerned with human language data. It contains helpful functions for tasks such as tokenization, cleaning, and topic modeling.

PyText [27] PyText is a deep-learning based NLP modeling framework built on PyTorch, providing pre-trained models for NLP tasks such as sequence tagging, classification, and contextual intent-slot models.

SpaCy[28] A remarkably fast Python library for modeling and processing text in 34 different languages. It includes pretrained models to predict named entities, part-of-speech tags and syntactic dependencies, as well as starter models designed for transfer learning. It also has tools for tokenization, text cleaning, and statistical modeling.

Stanford CoreNLP[29] A set of tools developed by Stanford NLP Group for statistical, neural, and rule-based problems in computational linguistics. Its software provides a simple, useful interface for NLP tasks like NER and part-of-speech (POS) tagging.

8.2.3 Clinical NLP. **Criteria2Query** [30] A system for automatically transforming clinical research eligibility criteria to Observational Medical Outcomes Partnership (OMOP) Common Data Model-based executable cohort queries. [258] The system is an information extraction pipeline that combines machine learning and rule-based methods.

CuiTools [31] A package of PERL programs for word sense disambiguation (WSD)[139]. Its models perform supervised or unsupervised WSD using both general English knowledge and specific medical concepts extracted from UMLS.

Metamap[32] A tool to identify medical concepts from the text and map them to standard terminologies in the UMLS. MetaMap uses a knowledge-intensive approach based on symbolic, NLP and computational-linguistic techniques.

MIMIC-Extract [33] An open source pipeline to preprocess and present data from MIMIC-III v1.4 database. [233] MIMIC-Extract has useful features for analysis - for example, it transforms discrete temporal data into a time-series, and extracts clinically relevant targets like mortality from the text.

ScispaCy[34] Many NLP models perform poorly under domain shift, so ScispaCy adapts SpaCy's models to process scientific, biomedical, or clinical text. It was developed by AllenNLP in 2019 and includes much of the same functionality as SpaCy.

[23] https://github.com/pytorch/fairseq
[24] https://fasttext.cc/
[25] https://radimrehurek.com/gensim/index.html
[26] https://www.nltk.org/
[27] https://pytext-pytext.readthedocs-hosted.com/en/latest/
[28] https://spacy.io/
[29] https://stanfordnlp.github.io/CoreNLP/
[30] http://www.ohdsi.org/web/criteria2query/
[31] http://cuitools.sourceforge.net/
[32] https://metamap.nlm.nih.gov/
[33] https://github.com/MLforHealth/MIMIC_Extract
[34] https://allenai.github.io/scispacy/

Unified Medical Language System (UMLS)[35] UMLS is a set of files and software that provides unifying relationships across a number of different medical vocabularies and standards. Its aim is to improve effectiveness and interoperability between biomedical information systems like EHRs. It can be used to link medical terms, drug names, or billing codes across different computer systems.

9 CONCLUSION AND FUTURE DIRECTION

In recent years, NLP has made rapid progress in the EHR domain. In this survey, we reviewed recent studies on how EHR tasks could benefit from deep NLP models. More specifically, we summarized the works for the following EHR-NLP tasks: classification and prediction, clinical embeddings, extraction, generation and summarization, and other topics including question answering, phenotyping, knowledge graphs, multilinguality and medical Dialogues. We also listed some relevant datasets and existing tools to promote EHR-NLP reseacrh.

Though structured data are mostly investigated for EHRs, in this survey, we mainly focused on understanding unstructured text data for downstream EHR tasks. One of the future direction maybe better mining knowledge and information from unstructured data [51], and a good combination both types for a better decision making. Besides, another direction could be utilizing transfer learning or unsupervised learning for EHR tasks, as usually very limited labeled data are available. We hope that this work will inspire the readers and promote NLP and EHR research.

REFERENCES

[1] Amjad Abu-Jbara and Dragomir R. Radev. 2011. Coherent Citation-Based Summarization of Scientific Papers. In *ACL 2011*, Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (Eds.). The Association for Computer Linguistics, Portland, Oregon, 500–509. https://www.aclweb.org/anthology/P11-1051/

[2] Griffin Adams, Ketenci Mert, Shreyas Bhave, Adler Perotte, and Elhadad Noémie. 2020. Zero-Shot Clinical Acronym Expansion via Latent Meaning Cells. https://arxiv.org/pdf/2010.02010

[3] Ahmad Al-Aiad, Rehab Duwairi, and Manar Fraihat. 2018. Survey: Deep Learning Concepts and Techniques for Electronic Health Record. In *15th IEEE/ACS International Conference on Computer Systems and Applications, AICCSA 2018, October 28 - Nov. 1, 2018*. IEEE Computer Society, Aqaba, Jordan, 1–55. https://doi.org/10.1109/AICCSA.2018.8612827

[4] Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. 2020. Exploring Transformer Text Generation for Medical Dataset Augmentation. In *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 4699–4708. https://www.aclweb.org/anthology/2020.lrec-1.578

[5] Emilia Apostolova, D. Channin, Dina Demner-Fushman, Jacob D. Furst, S. Lytinen, and D. Raicu. 2009. Automatic segmentation of clinical texts. , 5905–5908 pages. https://doi.org/10.1109/IEMBS.2009.5334831

[6] Michela Assale, L. G. Dui, Andrea Cina, Andrea Seveso, and F. Cabitza. 2019. The Revival of the Notes Field: Leveraging the Unstructured Content in Electronic Health Records. *Frontiers in Medicine* 6 (2019), 66.

[7] Pinkesh Badjatiya, Litton J. Kurisinkel, Manish Gupta, and Vasudeva Varma. 2018. Attention-based Neural Text Segmentation. *CoRR* abs/1808.09935 (2018), 180–193. arXiv:1808.09935 http://arxiv.org/abs/1808.09935

[8] Trapit Bansal, Patrick Verga, Neha Choudhary, and Andrew McCallum. 2019. Simultaneously Linking Entities and Extracting Relations from Biomedical Text Without Mention-level Supervision. arXiv:1912.01070 http://arxiv.org/abs/1912.01070

[9] Mrinal Kanti Baowaly, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. 2019. Synthesizing electronic health records using improved generative adversarial networks. *J Am Med Inform Assoc.* 26(3) (2019), 228–241. https://pubmed.ncbi.nlm.nih.gov/30535151/

[10] Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noʻemie Elhadad. 2017. Multi-Label Classification of Patient Notes a Case Study on ICD Code Assignment. arXiv:arXiv:1709.09587

[11] Andrew Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, G. Weber, Nathan Palmer, X. Shi, T. Cai, and I. Kohane. 2020. Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 25 (2020), 295 – 306.

[12] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *EMNLP-IJCNLP 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3613–3618. https://doi.org/10.18653/v1/D19-1371

[35] https://www.nlm.nih.gov/research/umls/

[13] Asma Ben Abacha and Dina Demner-Fushman. 2019. On the Summarization of Consumer Health Questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2228–2234. https://doi.org/10.18653/v1/P19-1215

[14] Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinformatics* 20, 1 (Oct 2019), 511.

[15] Asma Ben Abacha and Dina Demner-Fushman. 2019. A Question-Entailment Approach to Question Answering. arXiv:1901.08079 [cs.CL] https://arxiv.org/abs/1901.08079

[16] Michael Bendersky, Honglei Zhuang, Ji Ma, Shuguang Han, Keith Hall, and Ryan McDonald. 2020. RRF102: Meeting the TREC-COVID challenge with a 100+ runs ensemble.

[17] Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational linguistics* 22, 1 (1996), 39–71.

[18] Daniel Biś, Canlin Zhang, Xiuwen Liu, and Zhe He. 2018. Layered Multistep Bidirectional Long Short-Term Memory Networks for Biomedical Word Sense Disambiguation. , 313–320 pages.

[19] Kevin Brady and Afser Shariff. 2013. Virtual medical scribes: making electronic medical records work for you. *The Journal of medical practice management: MPM* 29, 2 (2013), 133.

[20] Rebecca F. Bruce and Janyce Wiebe. 1994. Word-Sense Disambiguation Using Decomposable Models. , 139–146 pages. https://doi.org/10.3115/981732.981752

[21] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. , 1721–1730 pages. https://doi.org/10.1145/2783258.2788613

[22] Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word Sense Disambiguation Improves Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, 33–40. https://www.aclweb.org/anthology/P07-1005

[23] Zhengping Che, Sanjay Purushotham, Robinder G. Khemani, and Yan Liu. 2015. Distilling Knowledge from Deep Networks with Applications to Healthcare Domain. arXiv:1512.03542 http://arxiv.org/abs/1512.03542

[24] Zhengping Che, Sanjay Purushotham, Robinder G. Khemani, and Yan Liu. 2016. Interpretable Deep Models for ICU Outcome Prediction. http://knowledge.amia.org/amia-63300-1.3360278/t004-1.3364525/f004-1.3364526/2500209-1.3364981/2493688-1.3364976

[25] Irene Y Chen, Shalmali Joshi, Marzyeh Ghassemi, and Rajesh Ranganath. 2020. Probabilistic Machine Learning for Healthcare.

[26] Qingyu Chen, Robert Leaman, Alexis Allot, Ling Luo, Chih-Hsuan Wei, Shankai Yan, and Zhiyong Lu. 2020. Artificial Intelligence (AI) in Action: Addressing the COVID-19 Pandemic with Natural Language Processing (NLP).

[27] Siyu Chen, Yingqi Jia, Zhiquan Liu, Huanhuan Shan, Mao Chen, Hao Yu, Liangxue Lai, and Zhanjun Li. 2020. Robustly improved base editing efficiency of Cpf1 base editor using optimized cytidine deaminases. *Cell discovery* 6, 1 (2020), 1–4.

[28] Po-Hsiang Chiu and George Hripcsak. 2017. EHR-based phenotyping: Bulk learning and evaluation. *Journal of Biomedical Informatics* 70 (June 2017), 35–51.

[29] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. , 1724–1734 pages. https://doi.org/10.3115/v1/d14-1179

[30] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. 2016. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. , 301–318 pages. http://proceedings.mlr.press/v56/Choi16.html

[31] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. 2016. RETAIN: Interpretable Predictive Model in Healthcare using Reverse Time Attention Mechanism. arXiv:1608.05745 http://arxiv.org/abs/1608.05745

[32] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. 2016. Multi-layer Representation Learning for Medical Concepts. , 1495–1504 pages. https://doi.org/10.1145/2939672.2939823

[33] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. 2017. GRAM: Graph-based Attention Model for Healthcare Representation Learning. , 787–795 pages. https://doi.org/10.1145/3097983.3098126

[34] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. 2018. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. arXiv:1703.06490 [cs.LG]

[35] Edward Choi, Cao Xiao, Walter Stewart, and Jimeng Sun. 2018. MiME: Multilevel Medical Embedding of Electronic Health Records for Predictive Healthcare. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). Curran Associates, Inc., Montreal, Canada, 4547–4557. http://papers.nips.cc/paper/7706-mime-multilevel-medical-embedding-of-electronic-health-records-for-predictive-healthcare.pdf

[36] Alexis CONNEAU and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc., Vancouver, Canada, 7059–7069. https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf

[37] HIT Consultant. 2015. Why unstructured data holds the key to intelligent healthcare systems [Internet]. Atlanta (GA): HIT Consultant; 2015. cited at 2019 Jan 15.
[38] Alister D Costa, Stefan Denkovski, Michal Malyska, Sae Young Moon, Brandon Rufino, Zhen Yang, Taylor Killian, and Marzyeh Ghassemi. 2020. Multiple Sclerosis Severity Classification From Clinical Text. arXiv:arXiv:2010.15316
[39] Martin R Cowie, Juuso I Blomster, Lesley H Curtis, Sylvie Duclaux, Ian Ford, Fleur Fritz, Samantha Goldman, Salim Janmohamed, Jörg Kreuzer, Mark Leenay, et al. 2017. Electronic health records to facilitate clinical research. *Clinical Research in Cardiology* 106, 1 (2017), 1–9.
[40] Sajad Darabi, Mohammad Kachuee, Shayan Fazeli, and Majid Sarrafzadeh. 2020. TAPER: Time-Aware Patient EHR Representation.
[41] Surabhi Datta and Kirk Roberts. 2020. Spatial Relation Extraction from Radiology Reports using Syntax-Aware Word Representations. *AMIA Summits on Translational Science Proceedings* 2020 (2020), 116.
[42] Spiros C Denaxas and Katherine I Morley. 2015. Big biomedical data and cardiovascular disease research: opportunities and challenges. *European Heart Journal-Quality of Care and Clinical Outcomes* 1, 1 (2015), 9–16.
[43] Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association* 24, 3 (2017), 596–606.
[44] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/n19-1423
[45] Dmitriy Dligach and Timothy A. Miller. 2018. Learning Patient Representations from Text. , 119–123 pages. https://doi.org/10.18653/v1/s18-2014
[46] Nan Du, Kai Chen, Anjuli Kannan, Linh Tran, Yuhui Chen, and Izhak Shafran. 2019. Extracting Symptoms and their Status from Clinical Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 915–925. https://doi.org/10.18653/v1/P19-1087
[47] Tracy Edinger, Dina Demner-Fushman, Aaron M Cohen, Steven Bedrick, and William Hersh. 2017. Evaluation of clinical text segmentation to facilitate cohort retrieval. In *AMIA Annual Symposium Proceedings*, Vol. 2017. American Medical Informatics Association, AMIA, Washington, DC, USA, 660.
[48] Aparna Elangovan, Melissa J. Davis, and Karin Verspoor. 2020. Assigning function to protein-protein interactions: a weakly supervised BioBERT based approach using PubMed abstracts. arXiv:2008.08727 https://arxiv.org/abs/2008.08727
[49] Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research* 22 (2004), 457–479.
[50] Jean-Baptiste Escudié, Alaa Saade, Alice Coucke, and Marc Lelarge. 2018. Deep representation for patient visits from electronic health records.
[51] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. 2019. A guide to deep learning in healthcare. *Nature Medicine* 25, 1 (Jan. 2019), 24–29. https://doi.org/10.1038/s41591-018-0316-z
[52] Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1074–1084. https://doi.org/10.18653/v1/P19-1102
[53] Richárd Farkas and György Szarvas. 2008. Automatic construction of rule-based ICD-9-CM coding systems. , S10 pages.
[54] José Luis Fernández-Alemán, Inmaculada Carrión Señor, Pedro Ángel Oliver Lozoya, and Ambrosio Toval. 2013. Security and privacy in electronic health records: A systematic literature review. *Journal of biomedical informatics* 46, 3 (2013), 541–562.
[55] Gregory P. Finley, Serguei V. S. Pakhomov, Reed McEwan, and Genevieve B. Melton. 2016. Towards Comprehensive Clinical Abbreviation Disambiguation Using Machine-Labeled Training Data. http://knowledge.amia.org/amia-63300-1.3360278/t004-1.3364525/f004-1.3364526/2500393-1.3364887/2498448-1.3364882
[56] Charles P Friedman, Adam K Wong, and David Blumenthal. 2010. Achieving a nationwide learning health system. *Science translational medicine* 2, 57 (2010), 57cm29–57cm29.
[57] Kunihiko Fukushima and Sei Miyake. 1982. Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognit.* 15, 6 (1982), 455–469. https://doi.org/10.1016/0031-3203(82)90024-3
[58] Joseph Futoma, Sanjay Hariharan, and Katherine A. Heller. 2017. Learning to Detect Sepsis with a Multitask Gaussian Process RNN Classifier. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017 (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, Sydney, Australia, 1174–1182. http://proceedings.mlr.press/v70/futoma17a.html
[59] Sajad Sotudeh Gharebagh, Nazli Goharian, and Ross W. Filice. 2020. Attend to Medical Ontologies: Content Selection for Clinical Abstractive Summarization. In *ACL 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for

Computational Linguistics, Online, 1899–1905. https://www.aclweb.org/anthology/2020.acl-main.172/
[60] Luka Gligic, Andrey Kormilitzin, Paul Goldberg, and Alejo J. Nevado-Holgado. 2020. Named entity recognition in electronic health records using transfer learning bootstrapped Neural Networks. *Neural Networks* 121 (2020), 132–139. https://doi.org/10.1016/j.neunet.2019.08.032
[61] Yoav Goldberg. 2016. A Primer on Neural Network Models for Natural Language Processing. *J. Artif. Intell. Res.* 57 (2016), 345–420. https://doi.org/10.1613/jair.4992
[62] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press, Cambridge, MA. http://www.deeplearningbook.org.
[63] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. arXiv:2007.15779 https://arxiv.org/abs/2007.15779
[64] Tracy D Gunter and Nicolas P Terry. 2005. The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions. *Journal of medical Internet research* 7, 1 (2005), e3.
[65] Itika Gupta, Barbara Di Eugenio, Brian D. Ziebart, Bing Liu, Ben Gerber, Lisa K. Sharp, Rafe Davis, and Aiswarya Baiju. 2018. Creating and Annotating a Corpus of Health Coaching Dialogue.
[66] Benedict Guzman, Isabel Metzger, Yindalon Aphinyanaphongs, and Himanshu Grover. 2020. Assessment of Amazon Comprehend Medical: Medication Information Extraction. arXiv:2002.00481 https://arxiv.org/abs/2002.00481
[67] Sadid A. Hasan, Bo Liu, Joey Liu, Ashequl Qadir, Kathy Lee, Vivek Datla, Aaditya Prakash, and Oladimeji Farri. 2016. Neural Clinical Paraphrase Generation with Attention. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*. The COLING 2016 Organizing Committee, Osaka, Japan, 42–53. https://www.aclweb.org/anthology/W16-4207
[68] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. PathVQA: 30000+ Questions for Medical Visual Question Answering.
[69] Jette Henderson, Joyce C. Ho, Abel N. Kho, Joshua C. Denny, Bradley A. Malin, Jimeng Sun, and Joydeep Ghosh. 2017. Granite: Diversified, Sparse Tensor Factorization for Electronic Health Record-Based Phenotyping.
[70] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.
[71] Norbert Hoffmann. 2013. *Simulation neuronaler Netze: Grundlagen, Modelle, Programme in Turbo Pascal*. Springer-Verlag, New York City, New York.
[72] A. Hoogi, A. Mishra, F. Gimenez, J. Dong, and D. Rubin. 2020. Natural Language Generation Model for Mammography Reports Simulation. *IEEE Journal of Biomedical and Health Informatics* 24, 9 (2020), 2711–2717.
[73] J J Hopfield. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences* 79, 8 (1982), 2554–2558. https://doi.org/10.1073/pnas.79.8.2554 arXiv:https://www.pnas.org/content/79/8/2554.full.pdf
[74] Chao-Chun Hsu, Shantanu Karnwal, Sendhil Mullainathan, Ziad Obermeyer, and Chenhao Tan. 2020. Characterizing the Value of Information in Medical Notes. arXiv:2010.03574 [cs.CL]
[75] Jinmiao Huang, Cesar Osorio, and Luke Wicent Sy. 2019. An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. *Computer Methods and Programs in Biomedicine* 177 (Aug 2019), 141–153. https://doi.org/10.1016/j.cmpb.2019.05.024
[76] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. arXiv:1904.05342 http://arxiv.org/abs/1904.05342
[77] Mark Hughes, I Li, Spyros Kotoulas, and Toyotaro Suzumura. 2017. Medical text classification using convolutional neural networks. *Stud Health Technol Inform* 235 (2017), 246–250.
[78] Dirk Hüske-Kraus. 2003. Text generation in clinical medicine–a review. *Methods of information in medicine* 42, 01 (2003), 51–60.
[79] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In *AAAI 2019*. AAAI Press, Honolulu, Hawaii, USA, 590–597. https://doi.org/10.1609/aaai.v33i01.3301590
[80] Abhyuday N Jagannatha and Hong Yu. 2016. 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*. Association for Computational Linguistics, San Diego, California, 473–482. https://doi.org/10.18653/v1/N16-1056
[81] Shivlu Jain. 2017. Neuron. http://www.mplsvpn.info/2017/11/what-is-neuron-and-artificial-neuron-in.html
[82] Stefan James, Sunil V Rao, and Christopher B Granger. 2015. Registry-based randomized clinical trials—a new clinical trial paradigm. *Nature Reviews Cardiology* 12, 5 (2015), 312–316.
[83] Min Jiang, Yukun Chen, Mei Liu, S Trent Rosenbloom, Subramani Mani, Joshua C Denny, and Hua Xu. 2011. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association* 18, 5 (2011), 601–606.

[84] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams.
[85] Qiao Jin, Bhuwan Dhingra, William W. Cohen, and Xinghua Lu. 2019. Probing Biomedical Embeddings from Language Models. arXiv:1904.02181 http://arxiv.org/abs/1904.02181
[86] Yonghao Jin, Fei Li, and Hong Yu. 2018. HYPE: A High Performing NLP System for Automatically Detecting Hypoglycemia Events from Electronic Health Record Notes.
[87] Karen Sparck Jones. 1994. Natural language processing: a historical review. , 3–16 pages.
[88] Venkata Joopudi, Bharath Dandala, and Murthy Devarakonda. 2018. A convolutional route to abbreviation disambiguation in clinical text. *Journal of biomedical informatics* 86 (2018), 71–78.
[89] Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cécile Paris, and C. Raina Macintyre. 2019. Survey of Text-Based Epidemic Intelligence: A Computational Linguistics Perspective. *ACM Comput. Surv.* 52, 6, Article 119 (Oct. 2019), 19 pages. https://doi.org/10.1145/3361141
[90] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, Mirella Lapata, Phil Blunsom, and Alexander Koller (Eds.). Association for Computational Linguistics, Valencia, Spain, 427–431. https://doi.org/10.18653/v1/e17-2068
[91] Kohei Kajiyama, Hiromasa Horiguchi, Takashi Okumura, Mizuki Morita, and Yoshinobu Kano. 2018. De-identifying Free Text of Japanese Dummy Electronic Health Records. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*. Association for Computational Linguistics, Brussels, Belgium, 65–70. https://doi.org/10.18653/v1/W18-5608
[92] Kaur Karus. 2019. Using Embeddings to Improve Text Segmentation.
[93] Manana Khachidze, Magda Tsintsadze, and Maia Archuadze. 2016. Natural language processing based instrument for classification of free text medical records.
[94] Faiza Khan Khattak, Serena Jeblee, Noah H. Crampton, Muhammad Mamdani, and Frank Rudzicz. 2019. AutoScribe: Extracting Clinically Pertinent Information from Patient-Clinician Dialogues. In *MEDINFO 2019: Health and Wellbeing e-Networks for All - Proceedings of the 17th World Congress on Medical and Health Informatics, Lyon, France, 25-30 August 2019 (Studies in Health Technology and Informatics, Vol. 264)*, Lucila Ohno-Machado and Brigitte Séroussi (Eds.). IOS Press, Lyon, France, 1512–1513. https://doi.org/10.3233/SHTI190510
[95] Hak Gu Kim, Yeoreum Choi, and Yong Man Ro. 2017. Modality-bridge transfer learning for medical image classification. , 5 pages.
[96] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1746–1751. https://doi.org/10.3115/v1/D14-1181
[97] Yea-Seul Kim, Jessica Hullman, Matthew Burgess, and Eytan Adar. 2016. SimpleScience: Lexical Simplification of Scientific Terminology. , 1066–1071 pages. https://doi.org/10.18653/v1/d16-1114
[98] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. http://arxiv.org/abs/1312.6114
[99] Sebastian Köhler, Leigh Carmody, Nicole Vasilevsky, Julius O B Jacobsen, Daniel Danis, Jean-Philippe Gourdine, Michael Gargano, Nomi L Harris, Nicolas Matentzoglu, Julie A McMurry, et al. 2019. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic acids research* 47, D1 (2019), D1018–D1027.
[100] Bevan Koopman, Guido Zuccon, Anthony Nguyen, Anton Bergheim, and Narelle Grayson. 2015. Automatic ICD-10 classification of cancers from free-text death certificates. *International Journal of Medical Informatics* 84, 11 (Nov. 2015), 956–965. https://doi.org/10.1016/j.ijmedinf.2015.08.004
[101] Mark A Kramer. 1991. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal* 37, 2 (1991), 233–243.
[102] Kundan Krishna, Sopan Khosla, Jeffrey P. Bigham, and Zachary C. Lipton. 2020. Generating SOAP Notes from Doctor-Patient Conversations. arXiv:2005.01795 https://arxiv.org/abs/2005.01795
[103] Kundan Krishna, Amy Pavel, Benjamin Schloss, Jeffrey P. Bigham, and Zachary C. Lipton. 2020. Extracting Structured Data from Physician-Patient Conversations By Predicting Noteworthy Utterances. arXiv:2007.07151 https://arxiv.org/abs/2007.07151
[104] Reno Kriz, Eleni Miltsakaki, Marianna Apidianaki, and Chris Callison-Burch. 2018. Simplification Using Paraphrases and Context-Based Lexical Substitution. , 207–217 pages. https://doi.org/10.18653/v1/n18-1019
[105] Clete A Kushida, Deborah A Nichols, Rik Jadrnicek, Ric Miller, James K Walsh, and Kara Griffin. 2012. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Medical care* 50, Suppl (2012), S82.
[106] Gloria Hyun-Jung Kwak and Pan Hui. 2019. DeepHealth: Deep Learning for Health Informatics. arXiv:1909.00384 http://arxiv.org/abs/1909.00384
[107] MDI Biological Laboratory and NC State University. 2020. Comparative Toxicogenomics Database. https://http://ctdbase.org/. Accessed: 2020-08-22.
[108] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. , 1188–1196 pages. http://proceedings.mlr.press/v32/le14.html

[109] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, Vol. 32. JMLR.org, Beijing, China, 1188–1196.

[110] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[111] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.* 36, 4 (2020), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

[112] Scott Lee. 2018. Natural Language Generation for Electronic Health Records. arXiv:1806.01353 http://arxiv.org/abs/1806.01353

[113] Yoong Keok Lee and Hwee Tou Ng. 2002. An Empirical Evaluation of Knowledge Sources and Learning Algorithms for Word Sense Disambiguation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*. Association for Computational Linguistics, Philadelphia, USA, 41–48. https://doi.org/10.3115/1118693.1118699

[114] Ivan Lerner, Jordan Jouffroy, Anita Burgun, and Antoine Neuraz. 2020. Learning the grammar of prescription: recurrent neural network grammars for medication information extraction in clinical texts. arXiv:2004.11622 [cs.CL]

[115] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, Online, 7871–7880. https://www.aclweb.org/anthology/2020.acl-main.703/

[116] Fei Li, Yonghao Jin, Weisong Liu, Bhanu Pratap Singh Rawat, Pengshan Cai, and Hong Yu. 2019. Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)–Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study. *JMIR Med Inform* 7, 3 (12 Sep 2019), e14830. https://doi.org/10.2196/14830

[117] Irene Li, Michihiro Yasunaga, Muhammed Yavuz Nuzumlalı, Cesar Caraballo, Shiwani Mahajan, Harlan Krumholz, and Dragomir Radev. 2019. A Neural Topic-Attention Model for Medical Term Abbreviation Disambiguation.

[118] Mingjie Li, Fuyu Wang, Xiaojun Chang, and Xiaodan Liang. 2020. Auxiliary Signal-Guided Knowledge Encoder-Decoder for Medical Report Generation. arXiv:2006.03744 https://arxiv.org/abs/2006.03744

[119] Ying Li, Sharon Lipsky Gorman, and Noemie Elhadad. 2010. Section classification in clinical notes using supervised hidden markov model. , 744–750 pages.

[120] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020. BeHRt: transformer for electronic Health Records. *Scientific Reports* 10, 1 (2020), 1–12.

[121] Salvador Lima, Naiara Perez, Montse Cuadros, and German Rigau. 2020. NUBES: A Corpus of Negation and Uncertainty in Spanish Clinical Texts.

[122] Anne Linton. 1990. QMR (Quick Medical Reference).

[123] Lucian Vlad Lita, Shipeng Yu, Stefan Niculescu, and Jinbo Bi. 2008. Large Scale Diagnostic Code Classification for Medical Patient Records. https://www.aclweb.org/anthology/I08-2125

[124] Hongfang Liu, Virginia Teller, and Carol Friedman. 2004. A multi-aspect comparison study of supervised word sense disambiguation. *Journal of the American Medical Informatics Association* 11, 4 (2004), 320–331.

[125] Jingshu Liu, Zachariah Zhang, and Narges Razavian. 2018. Deep EHR: Chronic Disease Prediction Using Medical Notes. arXiv:1808.04928 [cs.LG]

[126] Luchen Liu, Jianhao Shen, Ming Zhang, Zichang Wang, and Jian Tang. 2018. Learning the Joint Representation of Heterogeneous Temporal Events for Clinical Endpoint Prediction. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, Louisiana, US, 109–116. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17085

[127] Xiangan Liu, Keyang Xu, Pengtao Xie, and Eric P. Xing. 2018. Unsupervised Pseudo-Labeling for Extractive Summarization on Electronic Health Records. arXiv:1811.08040 http://arxiv.org/abs/1811.08040

[128] Leonardo Campillos Llanos, Catherine Thomas, Éric Bilinski, Pierre Zweigenbaum, and Sophie Rosset. 2020. Designing a virtual patient dialogue system based on terminology-rich resources: Challenges and evaluation. *Nat. Lang. Eng.* 26, 2 (2020), 183–220.

[129] Xinrui Lyu, Matthias Hüser, Stephanie L. Hyland, George Zerveas, and Gunnar Rätsch. 2018. Improving Clinical Predictions through Unsupervised Time Series Representation Learning. arXiv:1812.00490 http://arxiv.org/abs/1812.00490

[130] Sean MacAvaney, Arman Cohan, and Nazli Goharian. 2020. SLEDGE-Z: A Zero-Shot Baseline for COVID-19 Literature Search.

[131] Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish A. Talati, and Ross W. Filice. 2019. Ontology-Aware Clinical Abstractive Summarization. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, Paris, France, 1013–1016. https://doi.org/10.1145/3331184.3331319

[132] Diwakar Mahajan, Jennifer J. Liang, and Ching-Huei Tsou. 2020. Extracting Daily Dosage from Medication Instructions in EHRs: An Automated Approach and Lessons Learned. arXiv:2005.10899 [cs.CL]
[133] Igor Igor Mikhailovich Malioutov. 2006. *Minimum cut model for spoken lecture segmentation.* Ph.D. Dissertation. Massachusetts Institute of Technology.
[134] Ben J Marafino, Jason M Davies, Naomi S Bardach, Mitzi L Dean, and R Adams Dudley. 2014. N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. *Journal of the American Medical Informatics Association* 21, 5 (2014), 871–875.
[135] Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable Sentence Simplification. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 4689–4698. https://www.aclweb.org/anthology/2020.lrec-1.577
[136] Aurelie Mascio, Zeljko Kraljevic, Daniel Bean, Richard J. B. Dobson, Robert Stewart, Rebecca Bendayan, and Angus Roberts. 2020. Comparative Analysis of Text Classification Approaches in Electronic Health Records. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, BioNLP 2020, July 9, 2020*, Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii (Eds.). Association for Computational Linguistics, Online, 86–94. https://www.aclweb.org/anthology/2020.bionlp-1.9/
[137] Clara McCreery, Namit Katariya, Anitha Kannan, Manish Chablani, and Xavier Amatriain. 2019. Domain-Relevant Embeddings for Medical Question Similarity. arXiv:1910.04192 http://arxiv.org/abs/1910.04192
[138] Denis Jered McInerney, Borna Dabiri, Anne-Sophie Touret, Geoffrey Young, Jan-Willem van de Meent, and Byron C. Wallace. 2020. Query-Focused EHR Summarization to Aid Imaging Diagnosis. arXiv:2004.04645 https://arxiv.org/abs/2004.04645
[139] Bridget T McInnes, Ted Pedersen, and John Carlis. 2007. Using UMLS Concept Unique Identifiers (CUIs) for word sense disambiguation in the biomedical domain. , 533 pages.
[140] Saeed Mehrabi, Sunghwan Sohn, Dingcheng Li, Joshua J. Pankratz, Terry M. Therneau, Jennifer L. St. Sauver, Hongfang Liu, and Mathew J. Palakal. 2015. Temporal Pattern and Association Discovery of Diagnosis Codes Using Deep Learning. , 408–416 pages. https://doi.org/10.1109/ICHI.2015.58
[141] Oren Melamud and Chaitanya Shivade. 2019. Towards Automatic Generation of Shareable Synthetic Clinical Notes Using Neural Language Models. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 35–45. https://doi.org/10.18653/v1/W19-1905
[142] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (Lake Tahoe, Nevada) *(NIPS'13)*. Curran Associates Inc., Red Hook, NY, USA, 3111–3119.
[143] Randolph A. Miller, Harry E. Pople, and Jack D. Myers. 1982. Internist-I, an Experimental Computer-Based Diagnostic Consultant for General Internal Medicine. *New England Journal of Medicine* 307, 8 (1982), 468–476. https://doi.org/10.1056/NEJM198208193070803 arXiv:https://doi.org/10.1056/NEJM198208193070803 PMID: 7048091.
[144] Riccardo Miotto, Li Li, Brian A. Kidd, and Joel T. Dudley. 2016. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records.
[145] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. 2017. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics* 19, 6 (05 2017), 1236–1246.
[146] Rashmi Mishra, Jiantao Bian, Marcelo Fiszman, Charlene R Weir, Siddhartha Jonnalagadda, Javed Mostafa, and Guilherme Del Fiol. 2014. Text summarization in the biomedical domain: a systematic review of recent research. *Journal of biomedical informatics* 52 (2014), 457–467.
[147] Yasuhide Miura, Yuhao Zhang, Curtis P. Langlotz, and Dan Jurafsky. 2020. Improving Factual Completeness and Consistency of Image-to-Text Radiology Report Generation.
[148] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. 2018. *Foundations of machine learning.* MIT press, Cambridge, MA.
[149] Diego Mollá, Christopher Jones, and Vincent Nguyen. 2020. Query Focused Multi-document Summarisation of Biomedical Texts. arXiv:2008.11986 https://arxiv.org/abs/2008.11986
[150] Milad Moradi. 2019. Small-world networks for summarization of biomedical articles. arXiv:1903.02861 http://arxiv.org/abs/1903.02861
[151] Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How transferable are neural networks in nlp applications?
[152] Michael C. Mozer. 1989. A Focused Backpropagation Algorithm for Temporal Pattern Recognition. http://www.complex-systems.com/abstracts/v03_i04_a04.html
[153] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable Prediction of Medical Codes from Clinical Text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1101–1111. https://doi.org/10.18653/v1/N18-1100
[154] Andriy Mulyar and Bridget T. McInnes. 2020. MT-Clinical BERT: Scaling Clinical Information Extraction with Multitask Learning. arXiv:2004.10220 https://arxiv.org/abs/2004.10220

[155] Tsendsuren Munkhdalai and Hong Yu. 2017. Neural Semantic Encoders. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, 397–407. https://www.aclweb.org/anthology/E17-1038

[156] Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT press, Cambridge, MA.

[157] Vinod Nair and Geoffrey E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, Johannes Fürnkranz and Thorsten Joachims (Eds.). Omnipress, Haifa, Israel, 807–814. https://icml.cc/Conferences/2010/papers/432.pdf

[158] Daniel Neil, Michael Pfeiffer, and Shih-Chii Liu. 2016. Phased LSTM: Accelerating Recurrent Network Training for Long or Event-based Sequences. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). Annual Conference on Neural Information Processing Systems 2016, Barcelona, Spain, 3882–3890. http://papers.nips.cc/paper/6310-phased-lstm-accelerating-recurrent-network-training-for-long-or-event-based-sequences

[159] Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics* 9, 1 (2018), 12.

[160] Aitor García Pablos, Naiara Pérez, and Montse Cuadros. 2020. Sensitive Data Detection and Classification in Spanish Clinical Text: Experiments with BERT. In *LREC 2020*. European Language Resources Association, Marseille, France, 4486–4494. https://www.aclweb.org/anthology/2020.lrec-1.552/

[161] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2009), 1345–1359.

[162] Ellie Pavlick and Chris Callison-Burch. 2016. Simple PPDB: A Paraphrase Database for Simplification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Berlin, Germany, 143–148. https://doi.org/10.18653/v1/P16-2024

[163] Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019, Florence, Italy, August 1, 2019*, Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii (Eds.). Association for Computational Linguistics, Florence, Italy, 58–65. https://doi.org/10.18653/v1/w19-5006

[164] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[165] Naiara Pérez, Pablo Accuosto, Àlex Bravo, Montse Cuadros, Eva Martínez-Garcia, Horacio Saggion, and German Rigau. 2020. Cross-lingual semantic annotation of biomedical literature: experiments in Spanish and English. *Bioinform.* 36, 6 (2020), 1872–1880. https://doi.org/10.1093/bioinformatics/btz853

[166] Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. 2014. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association* 21, 2 (2014), 231–237.

[167] Ahmad Pesaranghader, Stan Matwin, Marina Sokolova, and Ali Pesaranghader. 2019. deepBioWSD: effective deep neural word sense disambiguation of biomedical text data. *Journal of the American Medical Informatics Association* 26, 5 (2019), 438–446.

[168] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, Marilyn A. Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, Louisiana, USA, 2227–2237. https://doi.org/10.18653/v1/n18-1202

[169] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. 2017. Predicting healthcare trajectories from medical records: A deep learning approach. *Journal of Biomedical Informatics* 69 (May 2017), 218–229.

[170] François Portet, Ehud Reiter, Jim Hunter, and Somayajulu Sripada. 2007. Automatic Generation of Textual Summaries from Neonatal Intensive Care Data. In *Artificial Intelligence in Medicine, 11th Conference on Artificial Intelligence in Medicine, AIME 2007, Amsterdam, The Netherlands, July 7-11, 2007, Proceedings (Lecture Notes in Computer Science, Vol. 4594)*, Riccardo Bellazzi, Ameen Abu-Hanna, and Jim Hunter (Eds.). Springer, Amsterdam, The Netherlands, 227–236.

[171] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[172] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 http://arxiv.org/abs/1910.10683

[173] Ganesh Ramakrishnan, Apurva Jadhav, Ashutosh Joshi, Soumen Chakrabarti, and Pushpak Bhattacharyya. 2003. Question Answering via Bayesian Inference on Lexical Relations. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*. Association for Computational Linguistics, Sapporo, Japan, 1–10. https://doi.org/10.3115/1119312.1119313

[174] Prem Ramaswami. 2015. A remedy for your health-related questions: health info in the Knowledge Graph. https://blog.google/products/search/health-info-knowledge-graph/
[175] Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen M. Voorhees, Lucy Lu Wang, and William R. Hersh. 2020. TREC-COVID: rationale and structure of an information retrieval shared task for COVID-19. *J. Am. Medical Informatics Assoc.* 27, 9 (2020), 1431–1436. https://doi.org/10.1093/jamia/ocaa091
[176] Anthony J. Robinson and F. Failside. 1987. Static and Dynamic Error Propagation Networks with Application to Speech Coding. , 632–641 pages. http://papers.nips.cc/paper/42-static-and-dynamic-error-propagation-networks-with-application-to-speech-coding
[177] Roland Roller, Madeleine Kittner, Dirk Weissenborn, and Ulf Leser. 2018. Cross-lingual Candidate Search for Biomedical Concept Normalization. arXiv:1805.01646 http://arxiv.org/abs/1805.01646
[178] Maya Rotmensch, Yoni Halpern, Abdulhakim Tlimat, Steven Horng, and David Sontag. 2017. Learning a health knowledge graph from electronic medical records. *Scientific reports* 7, 1 (2017), 1–11.
[179] Ignacio Rubio-López, Roberto Costumero, Héctor Ambit, Consuelo Gonzalo-Martín, Ernestina Menasalvas, and Alejandro Rodríguez González. 2017. Acronym Disambiguation in Spanish Electronic Health Narratives Using Machine Learning Techniques. *Studies in health technology and informatics* 235 (2017), 251—255.
[180] Sebastian Ruder. 2019. *Neural transfer learning for natural language processing*. Ph.D. Dissertation. National University of Ireland, Galway, Ireland.
[181] Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Florence, Italy, 15–18.
[182] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature* 323, 6088 (1986), 533–536.
[183] Stuart Russell and Peter Norvig. 2002. Artificial intelligence: a modern approach.
[184] Sunil Kumar Sahu and Ashish Anand. 2018. Drug-drug interaction extraction from biomedical texts using long short-term memory network. *J. Biomed. Informatics* 86 (2018), 15–24. https://doi.org/10.1016/j.jbi.2018.08.005
[185] Max E. Savery, Asma Ben Abacha, Soumya Gayen, and Dina Demner-Fushman. 2020. Question-Driven Summarization of Answers to Consumer Health Questions. https://arxiv.org/abs/2005.09067
[186] Martin Scaiano and Diana Inkpen. 2012. Getting More from Segmentation Evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Montréal, Canada, 362–366. https://www.aclweb.org/anthology/N12-1038
[187] Elyne Scheurwegs, Kim Luyckx, Léon Luyten, Walter Daelemans, and Tim Van den Bulcke. 2015. Data integration of structured and unstructured sources for assigning clinical codes to patient stays. *Journal of the American Medical Informatics Association* 23, e1 (08 2015), e11–e19. https://doi.org/10.1093/jamia/ocv115 arXiv:https://academic.oup.com/jamia/article-pdf/23/e1/e11/17377079/ocv115.pdf
[188] Allen Schmaltz and Andrew Beam. 2020. Exemplar Auditing for Multi-Label Biomedical Text Classification. arXiv:2004.03093 https://arxiv.org/abs/2004.03093
[189] Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural networks* 61 (2015), 85–117.
[190] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, Vancouver, Canada, 1073–1083. https://doi.org/10.18653/v1/P17-1099
[191] Sai P. Selvaraj and Sandeep Konam. 2019. Medication Regimen Extraction From Medical Conversations. arXiv:1912.04961 [cs.CL]
[192] Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P. Xing. 2017. Towards Automated ICD Coding Using Deep Learning. arXiv:arXiv:1711.04075
[193] Zhongmin Shi, Gabor Melli, Yang Wang, Yudong Liu, Baohua Gu, Mehdi M. Kashani, Anoop Sarkar, and Fred Popowich. 2007. Question Answering Summarization of Multiple Biomedical Documents. In *Advances in Artificial Intelligence, 20th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2007, Montreal, Canada, May 28-30, 2007, Proceedings (Lecture Notes in Computer Science, Vol. 4509)*, Ziad Kobti and Dan Wu (Eds.). Springer, Montreal, Canada, 284–295.
[194] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi. 2018. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE Journal of Biomedical and Health Informatics* 22, 5 (2018), 1589–1604.
[195] Han-Chin Shing, Guoli Wang, and Philip Resnik. 2019. Assigning Medical Codes at the Encounter Level by Paying Attention to Documents. arXiv:1911.06848 http://arxiv.org/abs/1911.06848
[196] M. Shwe, B. Middleton, D. Heckerman, M. Henrion, E. Horvitz, H. Lehmann, and G. Cooper. 1991. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. I. The probabilistic model and inference algorithms. *Methods of information in medicine* 30 4 (1991), 241–55.
[197] Stefano Silvestri, Francesco Gargiulo, Mario Ciampi, and Giuseppe De Pietro. 2020. Exploit Multilingual Language Model at Scale for ICD-10 Clinical Text Classification. In *IEEE Symposium on Computers and Communications, ISCC 2020, Rennes,France, July 7-10, 2020*.

IEEE, Rennes,France, 1–7. https://doi.org/10.1109/ISCC50000.2020.9219640
[198] A. K. Bhavani Singh, Mounika Guntu, Ananth Reddy Bhimireddy, Judy W. Gichoya, and Saptarshi Purkayastha. 2020. Multi-label natural language processing to identify diagnosis and procedure codes from MIMIC-III inpatient notes. arXiv:2003.07507 https://arxiv.org/abs/2003.07507
[199] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. 2020. CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. arXiv:2004.09167 https://arxiv.org/abs/2004.09167
[200] Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. 2017. BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics* 33, 14 (2017), i49–i58.
[201] Sarvesh Soni and Kirk Roberts. 2019. A Paraphrase Generation System for EHR Question Answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics, Florence, Italy, 20–29. https://doi.org/10.18653/v1/W19-5003
[202] Sarvesh Soni and Kirk Roberts. 2020. Paraphrasing to improve the performance of Electronic Health Records Question Answering. *AMIA Summits on Translational Science Proceedings* 2020 (2020), 626.
[203] Rafael T. Sousa, Lucas A. Pereira, and Anderson S. Soares. 2020. Predicting Diabetes Disease Evolution Using Financial Records and Recurrent Neural Networks. arXiv:1811.09350 [cs.LG]
[204] Mark Stevenson, Yikun Guo, Abdulaziz Alamri, and Robert Gaizauskas. 2009. Disambiguation of Biomedical Abbreviations. In *Proceedings of the BioNLP 2009 Workshop*. Association for Computational Linguistics, Boulder, Colorado, 71–79. https://www.aclweb.org/anthology/W09-1309
[205] Amber Stubbs, Michele Filannino, and Özlem Uzuner. 2017. De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID Shared Tasks Track 1. *Journal of biomedical informatics* 75 (2017), S4–S18.
[206] Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. *Journal of biomedical informatics* 58 (2015), S11–S19.
[207] Harini Suresh, Nathan Hunt, Alistair E. W. Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. 2017. Clinical Intervention Prediction and Understanding with Deep Neural Networks. In *Proceedings of the Machine Learning for Health Care Conference, MLHC 2017, Boston, Massachusetts, USA, 18-19 August 2017 (Proceedings of Machine Learning Research, Vol. 68)*, Finale Doshi-Velez, Jim Fackler, David C. Kale, Rajesh Ranganath, Byron C. Wallace, and Jenna Wiens (Eds.). PMLR, Massachusetts, USA, 322–337. http://proceedings.mlr.press/v68/suresh17a.html
[208] Madhumita Sushil, Simon Šuster, Kim Luyckx, and Walter Daelemans. 2018. Patient representation learning and interpretable evaluation using clinical notes. *Journal of Biomedical Informatics* 84 (Aug. 2018), 103–113.
[209] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., Montreal, Canada, 3104–3112. http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf
[210] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. 2018. A survey on deep transfer learning. , 270–279 pages.
[211] Buzhou Tang, Dehuan Jiang, Qingcai Chen, Xiaolong Wang, Jun Yan, and Ying Shen. 2019. De-identification of Clinical Text via Bi-LSTM-CRF with Neural Language Models. http://knowledge.amia.org/69862-amia-1.4570936/t004-1.4574923/t004-1.4574924/3203046-1.4574964/3201562-1.4574961
[212] Yifeng Tao, Bruno Godefroy, Guillaume Genthial, and Christopher Potts. 2018. Effective Feature Representation for Clinical Text Concept Extraction. arXiv:1811.00070 [cs.CL]
[213] Michael Tepper, Daniel Capurro, Fei Xia, Lucy Vanderwende, and Meliha Yetisgen-Yildiz. 2012. Statistical Section Segmentation in Free-Text Clinical Records. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey, 2001–2008. http://www.lrec-conf.org/proceedings/lrec2012/pdf/1016_Paper.pdf
[214] Jan Trienes, Dolf Trieschnigg, Christin Seifert, and Djoerd Hiemstra. 2020. Comparing Rule-based, Feature-based and Deep Neural Methods for De-identification of Dutch Medical Records. In *Proceedings of the ACM WSDM 2020 Health Search and Data Mining Workshop, at WSDM 2020 (CEUR Workshop Proceedings, Vol. 2551)*, Carsten Eickhoff, Yubin Kim, and Ryen W. White (Eds.). CEUR-WS.org, Texas, USA, 3–11. http://ceur-ws.org/Vol-2551/paper-03.pdf
[215] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics* 16, 1 (2015), 138.
[216] Alan Turing. 1950. Computing Machinery and Intelligence. *Mind* LIX, 236 (10 1950), 433–460. https://doi.org/10.1093/mind/LIX.236.433 arXiv:https://academic.oup.com/mind/article-pdf/LIX/236/433/30123314/lix-236-433.pdf
[217] Princeton University. 2020. WordNet A Lexical Database for English. https://wordnet.princeton.edu/. Accessed: 2020-08-22.
[218] Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association* 14, 5 (2007), 550–563.

[219] Ilya Valmianski, Caleb Goodwin, Ian M. Finn, Naqi Khan, and Daniel S. Zisook. 2019. Evaluating robustness of language models for chief complaint extraction from patient-generated text. arXiv:1911.06915 [cs.CL]

[220] Shikhar Vashishth, Rishabh Joshi, Ritam Dutt, Denis Newman-Griffis, and Carolyn Penstein Rosé. 2020. MedType: Improving Medical Entity Linking with Semantic Type Prediction. arXiv:2005.00460 https://arxiv.org/abs/2005.00460

[221] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need

[222] Alfredo Vellido. 2019. The importance of interpretability and visualization in machine learning for applications in medicine and health care. , 15 pages.

[223] David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A Healthcare Dataset for Complex Reasoning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 960–966. https://doi.org/10.18653/v1/p19-1092

[224] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008 (ACM International Conference Proceeding Series, Vol. 307)*, William W. Cohen, Andrew McCallum, and Sam T. Roweis (Eds.). ACM, Helsinki, Finland, 1096–1103. https://doi.org/10.1145/1390156.1390294

[225] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. Mach. Learn. Res.* 11 (2010), 3371–3408. http://portal.acm.org/citation.cfm?id=1953039

[226] Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2020. TREC-COVID: Constructing a Pandemic Information Retrieval Test Collection.

[227] Tu Vu, Baotian Hu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Sentence Simplification with Memory-Augmented Neural Networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, Marilyn A. Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, Louisiana, USA, 79–85. https://doi.org/10.18653/v1/n18-2013

[228] Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A Label Attention Model for ICD Coding from Clinical Text. https://doi.org/10.24963/ijcai.2020/461

[229] Ramya Vunikili, Supriya H. N, Vasile George Marica, and Oladimeji Farri. 2020. Clinical NER using Spanish BERT Embeddings. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), September 23th, 2020 (CEUR Workshop Proceedings, Vol. 2664)*. CEUR-WS.org, Málaga, Spain, 505–511.

[230] Alexander H. Waibel, Toshiyuki Hanazawa, Geoffrey E. Hinton, Kiyohiro Shikano, and Kevin J. Lang. 1989. Phoneme recognition using time-delay neural networks. *IEEE Trans. Acoust. Speech Signal Process.* 37, 3 (1989), 328–339. https://doi.org/10.1109/29.21701

[231] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The Covid-19 Open Research Dataset. arXiv:2004.10706 https://arxiv.org/abs/2004.10706

[232] Qingyun Wang, Manling Li, Xuan Wang, Nikolaus Parulian, Guangxing Han, Jiawei Ma, Jingxuan Tu, Ying Lin, Haoran Zhang, Weili Liu, et al. 2020. Covid-19 literature knowledge graph construction and drug repurposing report generation.

[233] Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. 2020. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. , 222–235 pages.

[234] Xuan Wang, Xiangchen Song, Yingjun Guan, Bangzheng Li, and Jiawei Han. 2020. Comprehensive named entity recognition on cord-19 with distant or weak supervision.

[235] Yanshan Wang, Sunghwan Sohn, Sijia Liu, Feichen Shen, Liwei Wang, Elizabeth J Atkinson, Shreyasee Amin, and Hongfang Liu. 2019. A clinical text classification paradigm using weak supervision and deep representation. *BMC medical informatics and decision making* 19, 1 (2019), 1.

[236] Yue Wang, Kai Zheng, Hua Xu, and Qiaozhu Mei. 2018. Interactive medical word sense disambiguation through informed learning. *Journal of the American Medical Informatics Association* 25, 7 (2018), 800–808.

[237] Zhenghui Wang, Yanru Qu, Liheng Chen, Jian Shen, Weinan Zhang, Shaodian Zhang, Yimei Gao, Gen Gu, Ken Chen, and Yong Yu. 2018. Label-Aware Double Transfer Learning for Cross-Specialty Medical Named Entity Recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, Marilyn A. Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, Louisiana, USA, 1–15. https://doi.org/10.18653/v1/n18-1001

[238] Xing Wei and Carsten Eickhoff. 2018. Embedding Electronic Health Records for Clinical Information Retrieval. arXiv:1811.05402 http://arxiv.org/abs/1811.05402

[239] Wei-Hung Weng, Yu-An Chung, and Peter Szolovits. 2019. Unsupervised Clinical Language Translation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis (Eds.). ACM, Anchorage, AK, USA, 3121–3131. https://doi.org/10.1145/3292500.3330710

[240] Paul Werbos. 1990. Backpropagation through time: what it does and how to do it. *Proc. IEEE* 78 (11 1990), 1550 – 1560. https://doi.org/10.1109/5.58337

[241] Paul J. Werbos. 1988. Generalization of backpropagation with application to a recurrent gas market model. *Neural Networks* 1, 4 (1988), 339–356. https://doi.org/10.1016/0893-6080(88)90007-X

[242] Ronald J. Williams. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Mach. Learn.* 8 (1992), 229–256. https://doi.org/10.1007/BF00992696

[243] Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, Bo Zhao, and Hua Xu. 2019. Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association* 27, 3 (12 2019), 457–470. https://doi.org/10.1093/jamia/ocz200 arXiv:https://academic.oup.com/jamia/article-pdf/27/3/457/32500129/ocz200.pdf

[244] Yonghui Wu, Jun Xu, Yaoyun Zhang, and Hua Xu. 2015. Clinical Abbreviation Disambiguation Using Neural Word Embeddings. In *Proceedings of BioNLP 15*. Association for Computational Linguistics, Beijing, China, 171–176. https://doi.org/10.18653/v1/W15-3822

[245] Yonghui Wu, Jun Xu, Yaoyun Zhang, and Hua Xu. 2015. Clinical Abbreviation Disambiguation Using Neural Word Embeddings. In *Proceedings of BioNLP 15*. Association for Computational Linguistics, Beijing, China, 171–176. https://doi.org/10.18653/v1/W15-3822

[246] Yonghui Wu, Xi Yang, Jiang Bian, Yi Guo, Hua Xu, and William Hogan. 2018. Combine factual medical knowledge and distributed word representation to improve clinical named entity recognition. , 1110 pages.

[247] Cao Xiao, Edward Choi, and Jimeng Sun. 2018. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association* 25, 10 (2018), 1419–1428.

[248] Hua Xu, Marianthi Markatou, Rositsa Dimova, Hongfang Liu, and Carol Friedman. 2006. Machine learning and word sense disambiguation in the biomedical domain: design and evaluation issues. *BMC bioinformatics* 7, 1 (2006), 334.

[249] Hua Xu, Peter D Stetson, and Carol Friedman. 2009. Methods for building sense inventories of abbreviations in clinical notes. *Journal of the American Medical Informatics Association* 16, 1 (2009), 103–108.

[250] Keyang Xu, Mike Lam, Jingzhi Pang, Xin Gao, Charlotte Band, Piyush Mathur, Frank Papay, Ashish K. Khanna, Jacek B. Cywinski, Kamal Maheshwari, Pengtao Xie, and Eric P. Xing. 2019. Multimodal Machine Learning for Automated ICD Coding. In *Proceedings of Machine Learning Research*, Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens (Eds.), Vol. 106. PMLR, Ann Arbor, Michigan, 197–215. http://proceedings.mlr.press/v106/xu19a.html

[251] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Trans. Assoc. Comput. Linguistics* 4 (2016), 401–415. https://transacl.org/ojs/index.php/tacl/article/view/741

[252] Shweta Yadav, Srivatsa Ramesh, Sriparna Saha, and Asif Ekbal. 2020. Relation Extraction from Biomedical and Clinical Text: Unified Multitask Learning Framework. arXiv:2009.09509 https://arxiv.org/abs/2009.09509

[253] Xi Yang, Tianchen Lyu, Qian Li, Chih-Yin Lee, Jiang Bian, William R Hogan, and Yonghui Wu. 2019. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Medical Informatics and Decision Making* 19, 5 (2019), 232.

[254] Liang Yao, Chengsheng Mao, and Yuan Luo. 2018. Clinical Text Classification with Rule-based Features and Knowledge-guided Convolutional Neural Networks. , 70–71 pages.

[255] Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R. Fabbri, Irene Li, Dan Friedman, and Dragomir R. Radev. 2019. ScisummNet: A Large Annotated Corpus and Content-Impact Models for Scientific Paper Summarization with Citation Networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, Hawaii, USA, 7386–7393. https://doi.org/10.1609/aaai.v33i01.33017386

[256] Wen-wai Yim, Meliha Yetisgen, Jenny Huang, and Micah Grossman. 2020. Alignment Annotation for Clinic Visit Dialogue to Clinical Note Sentence Language Generation. In *LREC 2020*. European Language Resources Association, Marseille, France, 413–421. https://www.aclweb.org/anthology/2020.lrec-1.52/

[257] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (Eds.). Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, Montreal, Canada, 3320–3328. http://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-neural-networks

[258] Chi Yuan, Patrick B Ryan, Casey Ta, Yixuan Guo, Ziran Li, Jill Hardin, Rupa Makadia, Peng Jin, Ning Shang, Tian Kang, et al. 2019. Criteria2Query: a natural language interface to clinical databases for cohort definition. *Journal of the American Medical Informatics*

Association 26, 4 (2019), 294–305.

[259] Klaus Zechner. 2002. Automatic Summarization of Open-Domain Multiparty Dialogues in Diverse Genres. *Computational Linguistics* 28, 4 (2002), 447–485. https://doi.org/10.1162/089120102762671945

[260] Z. Zeng, Y. Deng, X. Li, T. Naumann, and Y. Luo. 2019. Natural Language Processing for EHR-Based Computational Phenotyping. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 16, 1 (2019), 139–153.

[261] Canlin Zhang, Daniel Biś, Xiuwen Liu, and Zhe He. 2019. Biomedical word sense disambiguation with bidirectional long short-term memory and attention-based neural networks. *BMC bioinformatics* 20, 16 (2019), 502.

[262] Jinghe Zhang, Kamran Kowsari, James H. Harrison, Jennifer M. Lobo, and Laura E. Barnes. 2018. Patient2Vec: A Personalized Interpretable Deep Representation of the Longitudinal Electronic Health Record. *IEEE Access* 6 (2018), 65333–65346. https://doi.org/10.1109/ACCESS.2018.2875677

[263] Jingqing Zhang, Xiaoyu Zhang, K. Sun, Xian Yang, Chengliang Dai, and Y. Guo. 2019. Unsupervised Annotation of Phenotypic Abnormalities via Semantic Latent Representations on Electronic Health Records. , 598-603 pages.

[264] Quan-shi Zhang and Song-Chun Zhu. 2018. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering* 19, 1 (2018), 27–39.

[265] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). Curran Associates, Inc., Montreal, Canada, 649–657. http://papers.nips.cc/paper/5782-character-level-convolutional-networks-for-text-classification.pdf

[266] Xi Sheryl Zhang, Fengyi Tang, Hiroko H. Dodge, Jiayu Zhou, and Fei Wang. 2019. MetaPred: Meta-Learning for Clinical Risk Prediction with Limited Patient Electronic Health Records. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2487–2495. https://doi.org/10.1145/3292500.3330779

[267] Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. Learning to Summarize Radiology Findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis, Louhi@EMNLP 2018, Brussels, Belgium, October 31, 2018*, Alberto Lavelli, Anne-Lyse Minard, and Fabio Rinaldi (Eds.). Association for Computational Linguistics, Brussels, Belgium, 204–213. https://doi.org/10.18653/v1/w18-5623

[268] Yuhao Zhang, Derek Merck, Emily Bao Tsai, Christopher D. Manning, and Curtis Langlotz. 2020. Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, Online, 5108–5120. https://www.aclweb.org/anthology/2020.acl-main.458/

[269] Jiaping Zheng, Jorge Yarzebski, Balaji Polepalli Ramesh, Robert J Goldberg, and Hong Yu. 2014. Automatically detecting acute myocardial infarction events from EHR text: a preliminary study. , 1286 pages.

[270] Guoqiang Zhong, Li-Na Wang, Xiao Ling, and Junyu Dong. 2016. An overview on data representation learning: From traditional feature learning to recent deep learning. *The Journal of Finance and Data Science* 2, 4 (2016), 265–278.

[271] Zhi Zhong and Hwee Tou Ng. 2012. Word Sense Disambiguation Improves Information Retrieval. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 1: Long Papers*. The Association for Computer Linguistics, Jeju Island, Korea, 273–282. https://www.aclweb.org/anthology/P12-1029/

[272] Zihao Zhu, Changchang Yin, Buyue Qian, Yu Cheng, Jishang Wei, and Fei Wang. 2019. Measuring Patient Similarities via a Deep Architecture with Medical Concept Embedding. , arXiv–1902 pages.