Word Embeddings for Biomedical Natural Language Processing: A Survey

Billy Chiu
Language Technology Lab
University of Cambridge
billy1985322@gmail.com

Simon Baker
Language Technology Lab
University of Cambridge
sb895@cam.ac.uk

Abstract

Word representations are mathematical objects that capture the semantic and syntactic properties of words in a way that is interpretable by machines. Recently, encoding word properties into low-dimensional vector spaces using neural networks has become increasingly popular. Word embeddings are now used as the main input to Natural Language Processing (NLP) applications, achieving cutting-edge results. Nevertheless, most word-embedding studies are carried out with general-domain text and evaluation datasets, and their results do not necessarily apply to text from other domains (e.g. biomedicine) that are linguistically distinct from general English. To achieve maximum benefit when using word embeddings for biomedical NLP tasks, they need to be induced and evaluated using in-domain resources. Thus, it is essential to create a detailed review of biomedical embeddings that can be used as a reference for researchers to train in-domain models.

In this paper, we review biomedical word embedding studies from three key aspects: the corpora, models and evaluation methods. We first describe the characteristics of various biomedical corpora, and then compare popular embedding models. After that, we discuss different evaluation methods for biomedical embeddings. For each aspect, we summarize the various challenges discussed in the literature. Finally, we conclude the paper by proposing future directions that will help advance research into biomedical embeddings.

Keywords: word embeddings, biomedical NLP, evaluation

1 Introduction

Representation learning, when applied to textual data, generates word representations which capture the linguistic properties of words in a mathematical form (e.g. vectors). Each vector dimension corresponds to a feature that might have a semantic or syntactical interpretation (Turian et al., 2010). Most early studies employed human experts to propose a set of representative features for the data, which was expensive to obtain. Recently, an unsupervised approach, which encodes word meanings into a low-dimensional space using neural networks, has been proposed as an alternative (Bengio et al., 2003). Named *neural word representation* or *neural word embeddings*, this approach represents each word as a dense vector of real numbers, where synonyms appear as neighbors in the vector space. It can learn features in an unsupervised manner from large unlabeled corpora.

While word embeddings have been shown highly beneficial in recent works, most studies are carried out with general-domain text and evaluation datasets, and their results do not necessarily apply to text from other domains (e.g. biomedicine) that are linguistically distinct from general English. To achieve the maximum benefit when using word embeddings for biomedical Natural Language Processing (NLP) tasks, they need to be produced and evaluated using in-domain text. Thus, the goals of this article are to survey the cutting-edge models in vector space word embeddings and how they have been applied in biomedical text. We aim at covering topics for those who are new to the area, and also topics that give a new perspective to those who are already familiar with the area.

We assume our readers have a basic understanding of linear algebra (e.g., matrices and vectors). Additionally, we assume they have some familiarity with computational linguistics (e.g., vector space semantics) and deep learning (e.g., the Bidirectional Long Short-Term Memory, Bi-LSTM). However, if readers would like to do some further background reading, we refer them to Manning et al. (2008), Schütze and Pedersen (1993), as well as Hochreiter and Schmidhuber (1997) for further details.

We collected papers from various sources like PubMed, Google Scholar, ScienceDirect, ACL Web Anthology and AAAI. We confined to the papers which were mainly published in the period January 2016 to September 2020 because of the recent popularity of embeddings. We used keywords like "deep learning","biomedical", "clinical", "embeddings", "natural language processing" and "word representations" to retrieve the relevant papers and gathered 200 articles. After the removal of duplicate articles as well as the ones which were irrelevant to biomedical NLP, the number of articles were reduced to about 150. Finally, we focused on the most relevant 70 papers after a manual review of all the remaining articles.

This article is structured as follows. Section 2 provides a description of various cutting-edge embedding models. From Section 2.1 to 2.3, we present an overview and comparison of some embedding models (e.g., Word2vec) that are widely used in the general domain. After the high-level model framework is in place, Section 2.4 describes how these models have been applied in biomedical NLP tasks. In Section 3, we present a summary of various types of corpora for learning word embeddings. These include corpora of different domains (general English/biomedicine), sources (e.g., scientific literature and social media) and languages (English/non-English). By the end of Section 2 and 3, the readers will have a general understanding of various embedding models and their training corpora. We then take a detailed look at the evaluation methods for biomedical embeddings in Section 4. In particular, we mention two lines of evaluations: the intrinsic and extrinsic methods. The former measures the intrinsic properties of embedding models (e.g., how well it captures the notion of word similarity), whereas the latter measures how well individual models perform when used as features for extrinsic/downstream NLP tasks (e.g., relation classification). Based on the two lines of evaluations, we include some quantitative and qualitative performance from embeddings trained on difference corpora (biomedical v.s. non-biomedical) and model architectures. Later on, Section 5 considers the issues and directions of word embeddings, raising some questions about their power and their limitations. Finally, we conclude in Section 6.

2 Word Representation Models

The core principle of representation learning algorithms is developed on the basis of the distributional hypothesis, which suggests that lexical items with similar distributions share

similar meanings. More specifically, words that are used and occur in the same context tend to have similar meanings (Harris, 1954). Since distributional information (e.g. word co-occurrence counts) is largely available for many languages and can easily be extracted from large unannotated texts without depending on other NLP pipelines, the unsupervised learning of word representation using the distributional hypothesis has become widely popular.

In the literature, a wide range of representation models have been proposed. Early studies used bag-of-words models (e.g. word co-occurrence matrices, see Section 2.1) to represent the features of individual words. However, a high-dimensional, sparse matrix is needed to represent every word-word occurrence in corpora. There are several ways of reducing the dimensionality and sparsity of a matrix. Examples include Latent Semantic Analysis (LSA) (Deerwester et al., 1990; Foltz, 1996) which uses Singular Value Decomposition (SVD) (Golub and Reinsch, 1971) to decompose the co-occurrence matrix into a lower dimensional space, and Random Indexing (Kanerva et al., 2000) which considers the word-word occurrence within a fixed-size context window. Further, what is becoming particularly popular is encoding word semantics into a dense vector using a neural network, as proposed by Bengio et al. (2003). This method is commonly known as *Neural (word) embedding*. A neural embedding's learning algorithm functions much like a language modelling task, whose goal is to predict the next word given the previous ones in a sentence. Each word is represented as a finite-dimensional vector of real numbers, and the objective is to maximize the joint probability of a word and its context in terms of word vectors, using a feed-forward neural network. Word vectors are updated using back-propagation and gradient descent. Among these neural embeddings, the three widely-used models are the **Glo**bal **Ve**ctors for word representations (GloVe) by Pennington et al. (2014), the Word2vec by Mikolov et al. (2013a) and the FastText by Bojanowski et al. (2017) (see Section 2.1 and Section 2.2).

Word co-occurrences are frequently used to derive word representations since they are easy to obtain. For example, Levy and Goldberg (2014c) proposed a word representation approach based on SVD and a variant of the PMI (namely, Shifted Positive PMI). They showed that using the Shifted Positive PMI to represent words can obtain high performance when the corpus size is limited. Other word-word relations have also been considered. Regarding this, Levy and Goldberg (2014a) utilized the syntactic dependency relation to generate neural embeddings, better capturing the topicality of words. Further, Ammar et al. (2016) leveraged the word-pair mapping in parallel corpora to create multilingual embeddings. In particular, they used monolingual data, pairwise parallel dictionaries and clustering algorithms to induce neural embeddings of more than fifty languages. Additionally, Yu and Dredze (2014) proposed a new learning objective for modelling the word-word relations (synonyms) in resources other than corpora, such as lexicon or ontologies (WordNet) (Miller, 1995).

The aforementioned approaches make effective use of different word-word relations to derive embeddings for words, but they have several weaknesses that should be addressed. One arises when dealing with out-of-vocabulary (OOV) words: if a token has never been seen before, then the model does not have an embedding and it needs to fall back on a generic OOV representation. To tackle this, Luong et al. (2013) used a recursive neural network to model word features based on their morpheme composition. Additionally, Bojanowski et al. (2017) proposed FastText, which can handle OOV terms by extending the word2vec model with sub-word information, in the form of character n-grams (see Section 2.2.1). These types of character embeddings provide a better way of handling unseen words, whose representations can be constructed from vectors of known morphemes. Apart from

this, there are representation methods that leverage the unique structural properties existed in specific languages (e.g., strokes in Chinese). Regarding this, Cao et al. (2018) proposed to learn the representation of Chinese characters by considering their similarities in strokes formation. To illustrate, 口(mouth) and 目(eye) share similar strokes sequences: 丨, フ, 一, and such information can thus act as an indicator of their relatedness properties. Additionally, the similarities in radical components have been considered as features in learning Chinese and Japanese (Kanji) character representation (Chen et al., 2020; Toyama et al., 2017). The rationale is that similar characters share similar radicals. For example, 树(tree) and 森(forest) shares the radical: 木(wood).

For a single word, there can be different meanings depending on the context in text (i.e. word polysemy). Thus, it is ineffective to capture the word polysemy using only one embedding for each word. In view of this, Tian et al. (2014) proposed the Multi-prototype embeddings to represent different meanings of a word using multiple word embeddings, as derived from probabilistic models based on the different context of polysemous words. Recently, Peters et al. (2018) proposed **E**mbeddings from **L**anguage **Mo**dels (ELMo), which generate embeddings for words based on their context, by making use of the character embeddings and Bi-LSTM. For each word, the character embeddings encode its morpheme composition, and the Bi-LSTM encapsulates all preceding information, as well as all information that follows. This results in a highly contextual representation (see Section 2.3.1).

Representation models encode the semantic properties of words, and are used to provide features for NLP applications. These features can be learned in several ways. For example, Collobert and Weston (2008) proposed the learning of task-specific embeddings. Here, embeddings are learned as part of a neural network to solve a particular task (separate from word co-occurrence prediction). These embeddings capture task-specific word features (e.g., the nouns *Protein* and *Gene* are similar in Part-of-Speech (POS) tagging). Furthermore, Turian et al. (2010) demonstrated the usage of pre-trained neural embeddings in downstream applications such as named-entity recognition (NER) and chunking. Here, a word representation model is first learned from external corpora independent from the applications to be built (a.k.a. *pre-trained*). It is then adjusted (a.k.a. *fine-tuned*) to fit into an application as additional features. The idea of pre-training and fine-tuning embeddings based on the corresponding downstream tasks was incorporated into a relatively successful methodology, called **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (BERT) (see Section 2.3.2).

We have provided an overview of several representation models which encode word, character and contextual features. Among them, we pick five cutting-edge models (GloVe, Word2vec, FastText, ELMo and BERT) and describe them in detail, then we explain how they have been used in the biomedical domain.

2.1 GloVe

Global **Ve**ctors for word representations (GloVe) was proposed by Pennington et al. (2014). In GloVe, a word co-occurrence matrix is generated, where rows represent the words and columns represent the context. To illustrate this, consider the three sentences below:

1. I love chemistry.
2. I love maths.
3. I tolerate biology.

When co-occurrence frequencies are extracted at a sentence level, every word is said to be in the context of another word in the same sentence, and the corpus can be represented in the following matrix form:

$$\mathbf{M} = \begin{array}{c} \\ I \\ love \\ Chemistry \\ Maths \\ tolerate \\ Biology \end{array} \begin{array}{c} \begin{array}{cccccc} I & love & Chemistry & Maths & tolerate & Biology \end{array} \\ \begin{pmatrix} 0 & 2 & 1 & 1 & 1 & 1 \\ 2 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

M represents the matrix of the co-occurrence frequencies of words in the corpus. Each value in M is interpreted as how frequently a word co-occurs with its context. Factorization of the co-occurrence matrix results in a low-dimensional matrix, where rows represent words and columns represent features. Each row in the low-dimensional word-feature matrix is the dense vector representation of a word, where the size of the feature can be preset to the required value. The objective function in the GloVe model is

$$\sum_{i,j=1}^{V} f(X_{ij}(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - logX_{ij})^2 \tag{1}$$

Here, V refers to the vocabularies in the corpora and f is a weighting function. X_{ij} is the co-occurrence matrix for a target word (i) and its context word (j), and w_i, w_j, bi and b_j are a set of trainable parameters for i and j, where w_i and w_j are the embeddings, and b_i and b_j are their corresponding biases. Intuitively, GloVe learns neural embeddings by minimizing the reconstruction error between co-occurrence statistics predicted by the model and global co-occurrence statistics observed in the training corpus.

2.2 Continuous Bag-of-Words (CBOW) and Skip-gram

The CBOW and Skip-gram are two cutting-edge neural embedding algorithms introduced by Mikolov et al. (2013a,b) as part of the **Word2vec** tool. CBOW and Skip-gram have been shown to produce highly competitive neural embeddings in many intrinsic and extrinsic tasks (Pyysalo et al., 2013; Baker et al., 2016; Rei et al., 2016b; Tsvetkov et al., 2015), as compared to early models such as Random Indexing (Kanerva et al., 2000) and Latent Semantic Analysis (Landauer and Dumais, 1997), among others.

CBOW and Skip-gram learn neural embeddings through a neural network, which is composed of an input layer, a fully connected hidden layer, and an output layer. The size of the input layer is equal to the vocabulary size of the corpus (given a frequency filtering threshold), and each word is represented as a one-hot vector (i.e. a vector of size $|V|$ where one dimension is set to 1 to indicate a word, and other dimensions are set to 0). The hidden layer corresponds to the dimensions of the output word vectors. If a corpus consists of $|V|$ words whose word vectors are of dimension D, then the hidden layer will be a matrix of size $V \times D$, where each row corresponds to a word (as illustrated in Fig 1 and Fig 2). The output of the hidden layer is essentially the product of the hidden layer weight matrices (which are the learned embeddings). The size of the hidden layer is a hyper-parameter pre-defined by the users. While a higher dimension captures more word information, its training produces

a larger word representation matrix and is more computationally expensive (Mikolov et al., 2013c).

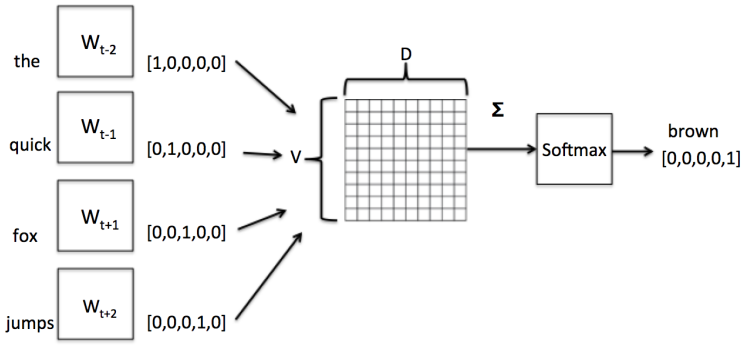

Figure 1: An illustration of the CBOW model with window size 2. The model is predicting the root word 'brown' given the context 'the quick fox jumps'. V is the total number of words in the corpus, and D is the dimension of the word vectors. The symbol $\sum$ indicates the average of the input context word ($c(w_t)$) vectors multiplied by the hidden layer weights. The Softmax function estimates a probability distribution over all words in the vocabulary.

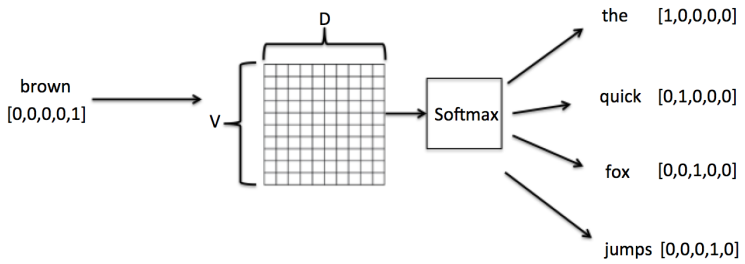

Figure 2: An illustration of the Skip-gram model with window size 2. The model is predicting the root word 'brown' given the context 'the'. Every word-context pair is trained individually. V is the total number of words in the corpus, and D is the dimension of the word vectors. The Softmax function estimates a probability distribution over all words in the vocabulary.

CBOW and Skip-gram aim to maximize the probability of an individual word given its context: $P(w_t|c(w_t)); w \in V$, where w_t refers to the root word (i.e. the target word to be trained), V is the vocabulary of the corpus, and $c(w_t)$ is the set of context words that surround the root. The size of the context window defines the range of words to be included as the context of a root word, which again, is a hyper-parameter pre-defined by users. For instance, a window size of 2 takes two words before and after a root word as its context for training. The window size is an important hyper-parameter in embedding learning models

because it controls the number of words to be considered as the context for representing an individual word. A wider window may be required when training on text that is full of long sentences containing complex clausal structures (e.g. biomedical literature). Additionally, it has been shown that the window size of a model influences the types of word semantics it captures: a larger window size emphasizes the learning of topic similarity between words, while a narrow context window leads the representation learning to primarily capture word functions (Turney, 2012).

A key difference between the trainings of CBOW and Skip-gram is their differentiated ways of denoting the context words (i.e. $c(w_t)$). In CBOW, context is defined as the average of word vectors $\vec{c_i}$ within the window (size = i), and is calculated as follows:

$$c(w_t) = \frac{1}{|c(w_t)|}\big(\sum_{c_i \in c(w_t)} \vec{c_i}\big)^\top \tag{2}$$

In contrast, Skip-gram considers each context word in a window as a distinct vector, which is calculated as:

$$c(w_t) = \big(\vec{c_i}\big)^\top \tag{3}$$

Consequently, the output layer generates a probability value for the root word. This is done by converting the activation values output by the hidden layer into probabilities using the Softmax function, as follows:

$$P(w_t|c(w_t)) = \frac{\exp(c(\vec{w_t})^\top \cdot \vec{w_t})}{\sum_{v_i \in V} \exp(c(\vec{w_t})^\top \cdot \vec{v_i})} \tag{4}$$

CBOW and Skip-gram operate on a *Log Bilinear* language model architecture. It computes the context vector $c(w_t)$ as a *linear combination* of the previous word vectors $\vec{c_i}$. The Log Bilinear refers to the part that the log of the numerator is a bilinear map for the context and root vectors (in Equation 4). The architectures of CBOW and Skip-gram share similar training parameters (e.g. context window size and vector dimension). Nevertheless, Skip-gram individually maps every word-context pair within a context window, making it intractable when used with a large amount of training data. Thus, its approximation counterpart – CBOW is introduced. This model only estimates the probability of each root word with the context average within the window. Other approximation techniques, such as negative sampling and sub-sampling, are also introduced as user-defined parameters in the Word2vec package. These parameters control the number of training samples and facilitate the effective Skip-gram training in a large corpus. However, it is still uncertain how these training parameters influence the quality of the learned model.

In general, both Word2vec and GloVe learn the representations of words based on the context. The difference between them is that Word2vec leverages intra-sentence (local) context, whereas GloVe incorporates matrix factorization methods in leveraging the inter-sentence (global) statistics of words.

2.2.1 FastText

Both GloVe and Word2vec are trained on word co-occurrence frequencies to capture word features at a word-level. However, many word formations, especially in biomedicine, follow morphological rules (e.g. *phosphorylate* and *dephosphorylate*). It is possible to improve embeddings by incorporating both word- and character-level information. With this in

mind, Bojanowski et al. (2017) proposed the FastText embedding model. FastText is an extension and improvement of the Word2vec model that incorporates sub-word information when learning embeddings. In the FastText model, each word is treated as a bag of character n-grams. Each character n-gram is mapped to a dense vector and the sum of these dense vectors represent the word. [1]

By utilizing sub-word information, FastText can provide embeddings for unseen words. This is because even if a testing word is unseen during training, its embedding can be obtained using the sum of its character n-grams.

2.3 Contextualized Embeddings

In GloVe, Word2vec and FastText, the features of each word is represented by a single vector, disregarding the fact that its meaning and sense vary according to the context in which it is used (i.e. polysemy). This isn't always desirable in many tasks where word sense ambiguity can lead to lower performance, such as parsing, semantic role labelling, and NER (Tenney et al., 2019). Contextualized embeddings solves this problem by generating a different embedding for the same word given different context as input. In this section we describe two important algorithms in this line of work: ELMo and BERT.

2.3.1 ELMo

Peters et al. (2018) introduced ELMo (**E**mbeddings from **L**anguage **Mo**dels) which generate contextual embeddings by considering the contexts and morphological structures of individual words at each state in text. This way, the embeddings of the same word can vary depending on their syntactical contexts and morphological structures in text. The model architecture of ELMo is shown in Figure 3. It consists of multiple layers of Bi-LSTMs with character-level embeddings to encode contextual and sub-word information. The character-level embeddings are generated by first converting each word to an appropriate representation using its character formation (see Figure 4). These character representations are then run through a convolutional layer, followed by a max-pool layer to get a fixed-length representation of the entire word. Using convolutional filters allows ELMo to capture the character n-gram features that lead to a more powerful representation. Additionally, similar to FastText, using character embeddings ensures that ELMo can form a valid representation for OOVs. Finally the character representation is passed through a two-layer highway network, which allows smoother information transfer through the input, before being fed into the Bi-LSTM layer. The vanilla ELMo has two layers of bidirectional LSTM, and the residual connection is added between the first and second layers. Residual connections are used to allow information to flow through a network directly, without passing through the (intermediate) non-linear activation layers. ELMo uses Bi-LSTM in training, so that its language model not only understands the next word, but also the previous word in the sentence. An embedding of each word (k) is obtained as a weighted sum of the character-level embeddings (X) and the hidden states of the Bi-LSTMs (H), as followed:

$$ELMo_k^{task} = \gamma_k \cdot (s_0^{task} \cdot X_k + s_1^{task} \cdot H_{1,k} + s_2^{task} \cdot H_{2,k}) \tag{5}$$

[1] For example, the vector for the word *neuron* is the sum of the vectors of *neur*, *euro*, *uron*, *neuro*, *euron* etc, if we pre-defined the hyper parameters **minn=4** and **maxn=5** which represent the sizes of smallest and largest character n-grams.

here, s_i are the weights on the word and hidden representations from the language model and γ_k is a task-specific scaling factor. This factor allows ELMo to fine-tune its representation based on the downstream tasks it is used for (e.g., document classification and sentiment analysis). To use ELMo in a task, the weights of the trained language model (i.e., $ELMo_k^{task}$) need to be frozen and concatenated with the input representation of the task-specific model. The weighting factors γ_k and s_i are then updated during training of the task-specific model.

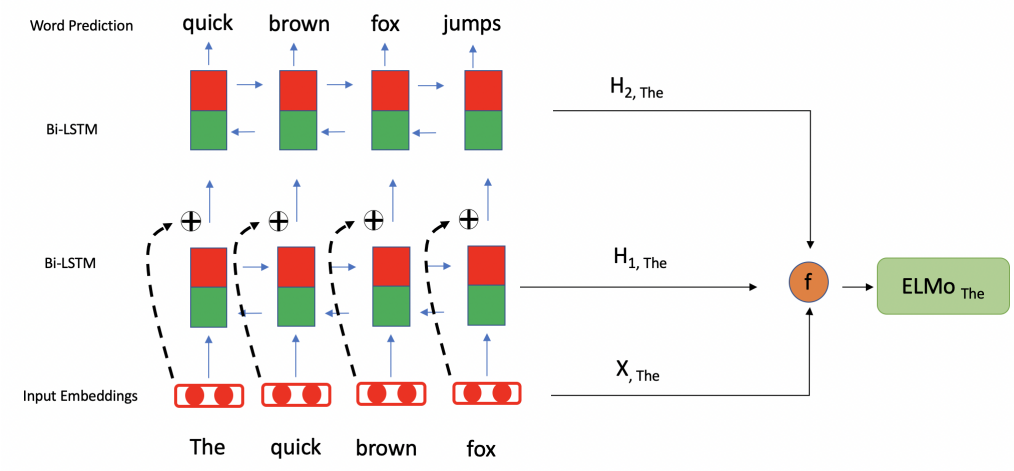

Figure 3: An illustration of the ELMo model. The model is predicting the root word 'The' given the context 'quick brown jumps'. X is the representation of an individual word which is learned based on its morpheme composition (see Figure 4), and H_1 and H_2 denote the hidden state representations of the two Bi-LSTM layers. The dotted line indicates a residual connection. The symbol f is a function that combines the three sets of representations (X, H_1 and H_2) into the final output.

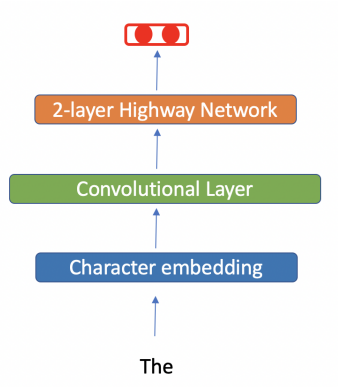

Figure 4: An illustration of how the word representation of 'The' is generated from its morpheme composition. Here, 'The' is first embedded as its individual character form: T, h, e, then the character embeddings run through the convolutional layer and highway network to get a fixed-length representation of the entire word.

2.3.2 BERT

Similar to ELMo, **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (BERT) (Devlin et al., 2019) is a contextualized neural embedding model, which learns embeddings by leveraging the contextual relationships between words (or sub-words) in text. The model architecture of BERT is shown in Figure 5. It consists of four parts: the input is a sequence of words, which are first embedded into vectors (*Input Embeddings*) and then processed by the *Transformer Encoder* (Vaswani et al., 2017). The output is a set of extracted feature embeddings (*Output Embeddings*), each of which represents local context for the corresponding words in the document. The output embeddings are used as features to produce *predictions* for two tasks: *Masked Language Modeling* (MLM) and *Next Sentence Prediction* (NSP). We will now describe each part in detail.

Input embeddings

In BERT, each input embeddings is a combination of three embeddings, namely:

1. **Token Embeddings**: The embeddings learned for the particular word from the *WordPiece* vocabulary (Yu et al., 2018).
2. **Sentence Embeddings**: In NSP, BERT predicts whether a given pair of sentences are consecutive. To help BERT distinguish between them, a sentence embedding is added to indicate, for a particular word, whether it belongs to sentence A or B.
3. **Position Embeddings**: These encode the position of each word in a sentence.

Transformer Encoder

A Transformer Encoder is composed of two main layers. The first is a (multi-head) **Self-Attention** layer which models the contextual features for individual words. The second is a position-wise fully connected feed-forward network to compute non-linear hierarchical features. Around both of these two layers, BERT employs a residual connection, followed by layer normalization (Ba et al., 2016) to stabilize the learning process. The output of each sub-layer is normalized as : $LayerNorm(x + Sublayer(x))$, where $Sublayer(x)$ is the function operated by the sub-layer itself.

A Self-Attention layer maps the input embeddings, a query (Q) and a set of key-value pairs (K, V) to an output vector using an *attention* mechanism (see Figure 6). Given the query and key vectors of dimension d_k, as well as value vectors of dimension d_v, the attention score is calculated as:

$$Attention(Q, K, V) = softmax(\frac{QK}{\sqrt{d_k}})V \tag{6}$$

Here, the layer computes the dot products of the query (Q) with all keys (K), divides each by $\frac{1}{\sqrt{d_k}}$, and applies a softmax function to obtain the weights on the value vectors. When the encoder is learning the embeddings for a word at a certain position, the attention score determines how much focus it should place on other parts of the input sentence. Instead of applying the attention function on a single set of Q, K, V, Vaswani et al. (2017) reported that it is beneficial to linearly project the queries, keys and values h times with different, learned linear projections and apply the attention function on all projections in parallel (a.k.a. multi-head attention). That way, the encoder can jointly attend to information from

different representation subspaces at different positions. In its base form, BERT includes 12 transformer encoders, 12 attention heads and 110M parameters for learning embeddings.

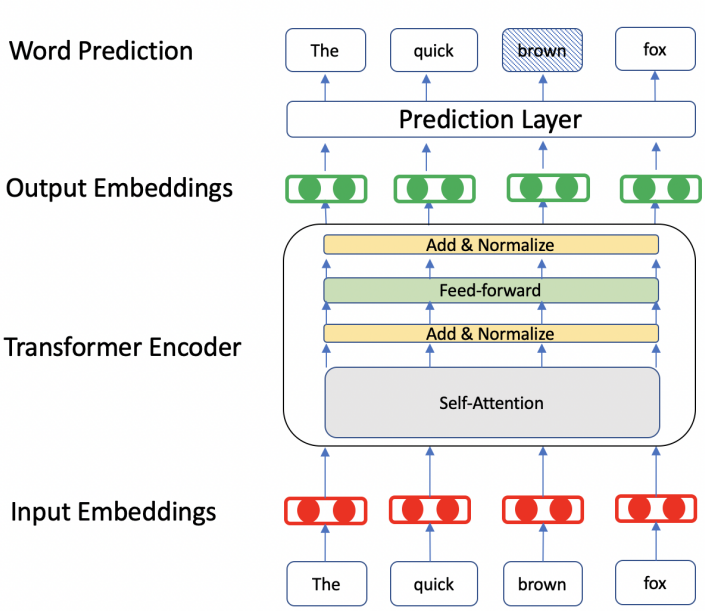

Figure 5: An illustration of the BERT model. The model is predicting the masked word 'brown' (shaded) given the context 'The quick fox'. The input text is first embedded into vectors (*Input Embeddings*) and then processed by the *Transformer Encoder* (details in Figure 6). The output is a set of extracted feature embeddings (*Output Embeddings*), which are then used as features for making a prediction (*Word Prediction*).

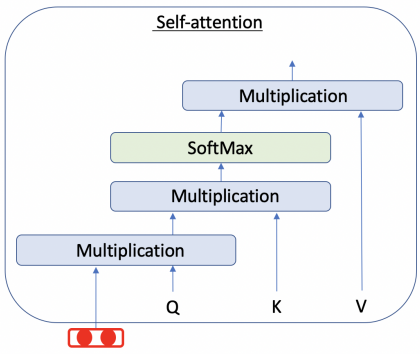

Figure 6: An illustration of the Self-Attention mechanism in the Transformer Encoder.

Output for Prediction

For language modelling, BERT trains the model parameters using two training objectives, masked language model (MLM) and next-sentence prediction (NSP). In MLM, BERT tries to predict a randomly masked word of the sequence using its context. In NSP, BERT predicts whether or not a given pair of sentences are consecutive in the document. We will now describe the two tasks.

When inputting word sequences to BERT, 15% of words in each sequence are first masked (i.e., shaded word in Figure 5). The model then attempts to predict them, based on the context provided by other non-masked words in the sequence. This way, words in different positions have roughly 'equal opportunity' to be trained. This differentiates BERT from other typical language models like ELMo, which look at a text sequence either from left-to-right or combined left-to-right and right-to-left training, resulting in locality bias (i.e., models focus on words that are closer to a target-trained word).

During training, BERT also learns to predict if the second sentence in a sentence pair is the subsequent one in the original document. Here, half of the training instances form pairs in which the second sentence is the subsequent sentence, and the other half are negative examples which are randomly picked and are disconnected from the first sentence.

For representation learning, BERT is trained on MLM and NSP tasks to understand the relations between words and sentences (respectively). Alternatively, the learned embeddings can also be used as features for other supervised NLP tasks (a.k.a., fine-tuning). Fine-tuning BERT for prediction tasks requires adding a prediction layer on top of the encoder output. The advantage of such an approach is that less parameters need to be learned from scratch. Several works have shown that BERT can be transferred for tasks, such as text summarization and sentiment analysis (Liu and Lapata, 2019; Li et al., 2019). Biomedical NLP researchers have also demonstrated the importance of transfer learning from pre-trained BERT, where the state-of-the-art performances are obtained by fine-tuning BERT with large task-/domain-specific data in NER, question answering (QA) and relation extraction (Lee et al., 2019).

We have described five cutting-edge embeddings (GloVe, Word2vec, FastText, ELMo and BERT). A summary is provided in Table 1. In general, word-level embeddings like Word2vec and GloVe are fast-to-train and easy-to-use (as compared with FastText, ELMo and BERT). These, however, come at a cost. First, these models ignore the morphological information and fail to handle OOVs like FastText does. Additionally, they do not account for the polysemous nature of words likes ELMo and BERT do. The latter models, however, are computationally and time expensive to train. Given these models, we will now describe how they have been used in the biomedical domain.

Models	Architecture	Strength	Weakness
CBOW	Log Bilinear	Fast to train on large corpora	Considerd the aggregate (average) of context instead of every word-context pair during training. Ignores morphological information
Skip-gram	Log Bilinear	Considers every word-context pair during training	Slower to train compared with CBOW Ignores morphological information
GloVe	Log Bilinear	Consider both local word-contexts pair, as well as global co-occurrence statistics	Ignores morphological information
FastText	Log Bilinear	Encodes morphological information	Computationally expensive
ELMo	Bi-LSTM	Considers both contextual and morphological information	Computationally expensive Encodes left and right contexts (individually) using separate LSTM
BERT	Transformer	Considers both contextual and morphological information Encodes left and right contexts (simultaneously) using MLM	Computationally expensive

Table 1: Summary of embedding models

2.4 Applications of Embedding Models

The linguistic properties of biomedical text differ significantly from general English (e.g, word usages, sentence length and writing styles). For examples, the verb *fire* can be used differently in the two domains (*fire a neuron* v.s. *fire a gun*). Additionally, biomedical named entities are often composed of long sequences of tokens (Leser and Hakenberg, 2005), and it is common to have alternate spellings and/or abbreviations for identical entities (Goulart et al., 2011). Furthermore, many biomedical terminologies are rarely found in general English dictionaries (Krauthammer and Nenadic, 2004). These make it difficult to directly use embeddings that are trained with general text for biomedical NLP. Hence, many researchers train their embeddings using large, openly available biomedical corpora, or fine-tune their embeddings with task-specific data as part of an end-to-end system.

Pyysalo et al. (2013) created the first set of neural embeddings in the biomedical domain using a collection of literature from PubMed and PMC. They generated embeddings using the skip-gram model. In addition, they created a word cluster representation by running k-means on the induced embeddings space, using word clusters to represent groups of similar words. Evaluation on three biomedical NER tasks showed that the skip-gram embeddings could provide useful features and a further clustering on top of the induced embeddings could lead to better results. However, they also mentioned that the skip-gram failed to capture the multiple-word representations, which led to a worse result in one of the NER tasks when it was compared with the multiple-word representations models from Stenetorp et al. (2012).

One drawback of word-level embeddings is that they cannot handle OOVs. If a token has never been seen before, then it does not have an embedding and the model needs to fall back on a generic OOV representation. This is a concern for languages with large vocabulary and many rare words or terminologies (e.g., biomedicine). For example, when classifying medical notes into standard disease codes, Karmakar (2018) reported a significantly high amount of OOVs, with the pre-trained GloVe embeddings covering less than half of the terminologies mentioned in the training corpora. As such, they decided to generate in-domain neural embeddings using the GloVe algorithm on the MIMIC III corpus.

They found that the in-domain embedding with the Convolutional Neural Network (CNN) (LeCun et al., 1995) provided the best performance for the task. Nowadays, character-level embeddings are commonly used in end-to-end bioinformatics applications to avoid the OOV issue. For example, Rei et al. (2016a) proposed an attention model for NER which combined the word-level embeddings and character-level embeddings using an attention mechanism. For word-level embeddings, they directly took the off-the-shelf embeddings that had been trained on Google News. The character-level embeddings were learned on-the-fly during the NER training. For evaluation, they also considered two alternatives: (1) the concatenation of character- and word-level embeddings, and (2) generic word-level embeddings. They evaluated their models on four biomedical NER datasets and found that using both the word- and character-level embeddings always gave a better performance, and an attention mechanism allowed the model to select which embeddings to use for each particular dataset. Apart from this, Le et al. (2018) proposed a novel CNN combined with multi-channel LSTM models for biomedical relation extraction tasks. Unlike in Rei et al. (2016a), where the character-level embeddings were induced during training, the authors pre-trained FastText embeddings on Wikipedia, and incorporated them as features for relation extraction tasks. In their experiment, they reported that the FastText embeddings provided the most important contribution to the model performance.

Embedding models like Word2Vec, GloVe, FastText assign a single vector representation to each word, independent of the context in which it is used. Nevertheless, using a single representation for a word cannot account for all its meanings in different contexts. To leverage the contextual information, Zhu et al. (2018) used the bi-LSTM CRF model, with contextual embeddings (ELMo) as input for NER. The ELMo embeddings were pre-trained on a corpus of medical-related Wikipedia pages, discharge summaries and radiology reports from MIMIC III. Experimental results showed that the domain-specific ELMo embeddings improved the performance of the model. Besides ELMo, BERT has also been applied in various subdomains in biomedicine (e.g., ClinicalBERT, BioBERT and SciBERT) and it continuously achieves cutting-edge performance in different tasks. Their details will be described in Sections 4.3 and 5.10.

2.4.1 Code Embeddings

EHRs describe patient information in the form of free text and medical codes. Standardized medical codes, such as the International Classification of Diseases (ICD-9, ICD-10), are widely used by doctors to ensure consistency in recording patient symptoms and diagnoses. Understanding the relations among different medical codes can contribute to bioinformatics tasks such as cohort selection and patient summarization. For example, doctors can effectively identify similar patients by looking up related ICD-9s in their reports. For this, embedding models have been used to capture the relatedness of medical codes.

Choi et al. (2016) applied a skip-gram on a proprietary dataset of medical claims to embed all the medical codes (diagnosis codes, procedure codes, laboratory codes, drug codes) into the same vector space. The dataset consists of diagnosis records for about four million patients from 2005 to 2013, including their diagnose codes (ICD-9), medical visits, lab test results, and drug usage in temporal sequence (see Figure 7). Instead of learning skip-gram embeddings from 'bag-of-words', the authors aimed to learn embeddings from 'bag-of-codes' in the claim data, where similar medical codes share similar contexts (i.e., neighboring codes). To measure the code relatedness as captured by the model, the authors proposed an evaluation method named the Medical Relatedness Measure (MRM).

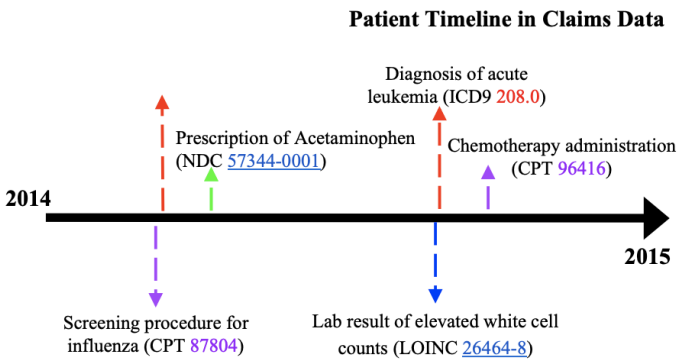

Figure 7: An illustration of the data used to learn code embeddings, sourced from Choi et al. (2016)

This compares the nearest neighbors of a certain set of medical codes in the embedding space with the hierarchical ICD-9 groupings from the Agency for Healthcare Research and CCS (details in Section 4.1.2). They reported that their code embeddings, as induced on the claim data, performed better than the one learned from textual data in the scientific abstracts (OHSUMED).

Typically, when training bag-of-codes embeddings, the model considers all code co-occurrences within a fixed-size window as indications of contexts, assuming every co-occurrence contributes equally to the embeddings training. However, as seen in Figure 7, medical codes are arranged in temporal sequence, which implies that, for a target code, the medical code that appears long ago may not be as relevant as the ones appearing just next to it. In view of this, Cai et al. (2018) proposed an attention CBOW model to learn time-aware code embeddings. Similar to Choi et al. (2016), their model was trained on bag-of-codes context from EHRs, except that they incorporated an attention mechanism on top of a CBOW model. That way, when learning the embeddings for a target code, the model also learned which contexts were more important and should be focused on. For evaluation, they also used MRM and measured the code relatedness using nearest neighbor search on the ICD-9 groupings by CCS. They reported better results with their time-aware code embeddings, as compared with the one from Choi et al. (2016) and other generic bag-of-words embeddings such as skip-gram, CBOW and GloVe.

Medical code features have been shown to be beneficial to downstream applications. For example, Che et al. (2017) proposed a CNN with pre-trained code embeddings for the task of risk prediction. To train the code embeddings, they first obtained proprietary EHR data, which contains the records for about 218k heart failure and diabetes patients as a sequence of medical events/codes. Then, they used the CBOW model with a context window of size 20 to generate the embeddings of size 200 for all the event codes with a minimum occurrence of five. The pre-trained embeddings were later used as features in a CNN model for predicting a given patient's risk of having diabetes or heart failure. They reported better results with the code embeddings, as compared with using random initialization or one-hot encoding features.

```
Bronchopulmonary Dysplasia was first described by Northway and colleagues in
1967 as a lung injury in a preterm infant resulting from oxygen and
mechanical ventilation.

C0006287 was first described by northway and colleagues in 1967 as a C0024109
C3263722 in a C0021294 C0678226 C0030054 and C0199470
```

Figure 8: An illustration of normalizing CUIs in text for learning CUI embeddings (e.g., Bronchopulmonary Dysplasia: C0006287). Sourced from Beam et al. (2020)

2.4.2 CUI Embeddings

UMLS is a unified database in biomedicine which integrates terminologies, classification and coding standards found in different resources. It groups all the synonyms from different vocabularies into a single concept and assigns it a unique identifier, called the Concept Unique Identifier (CUI). It follows a pattern of eight characters starting with C followed by seven digits (e.g., C0018681: headache). An accurate CUI representation can improve the efficiency of biomedical information retrieval. For example, a query search involving a specific UMLS term can be expanded to include related concepts, e.g., the 10 nearest neighbors in the embedded space. For this, embedding models have been used to capture the relatedness of CUIs.

De Vine et al. (2014) generated CUI embeddings by applying the skip-gram model on the OHSUMED dataset, which consists of about 348k medical journal abstracts. To train CUI embeddings, the authors first normalized the CUI entities in free text using MetaMap v11.2 (see Figure 8). After obtaining the CUI sequences, they learned the embeddings using a skip-gram, by representing individual CUIs using their contextual CUIs. They measured the CUI similarity captured by the embeddings by comparing their model output with the human gold standard using *MayoSRS* (a Word Similarity datasets, details in Section 4.1.1). They reported that using neural embeddings (i.e., skip-gram) to learn CUI embeddings yielded a better results as compared with other count-based methods like LSA and Random Indexing.

Instead of using journal abstracts, Choi et al. (2016) demonstrated how to learn CUI embeddings co-occurrence counts derived from clinical narratives. The data consists of the co-occurrence matrices of 1 million CUIs as extracted from the raw text of 20 million clinical notes from the Stanford Hospital and Clinics (Finlayson et al., 2014). To build CUI embeddings, the authors sampled word pairs proportional to the co-occurrence counts, then they presented these word pairs as training examples to the skip-gram. To measure the CUI similarity as captured by the model, the authors proposed an evaluation method named the Medical Conceptual Similarity Measure (MCSM). In particular, they compared the nearest neighbors of CUIs in the embedding space with the grouping of six medical concept types from the UMLS, namely *pharmacological substance, disease or syndrome, neoplastic process, clinical drug, finding*, and *injury or poisoning* (details in Section 4.1.2). They reported that skip-gram embeddings could better capture the CUI similarity from the co-occurrence matrix, as compared to other matrix factorization approaches like SVD.

Recently, Beam et al. (2020) presented a comprehensive set of CUI embeddings, namely *cui2vec*, learned using a large collection of biomedical data from different sources. This collection consists of the insurance claims of about 60 million members, 20 million clinical notes, and 1.7 million full text biomedical journal articles. The authors first used CUIs

to map/normalize all the medical codes, concepts and entities found in the data. Then, they trained the 'bag-of-CUIs' embeddings using the skip-gram and GloVe models. For evaluation, the authors included five word-similarity benchmarks to measure the embedding similarity for *co-morbidity, causative, drug-condition* and *synonymous* relations, as extracted from different sources, including the Mayo Clinic Encyclopedia of Diseases and Conditions, the National Drug File Reference Terminology and UMLS. Benefiting from the large amount of training data, cui2vec outperformed the cutting-edges models by Choi et al. (2016) and De Vine et al. (2014) on nearly all benchmarks.

2.4.3 Augmented Embeddings

Embedding models typically capture word semantics based on the distribution information from corpora (e.g., word co-occurrences). Indeed, word semantics can also be found in other data sources, such as terms in UMLS or MeSH hierarchy. These concepts can be integrated into the models using several methods.

One type of method involves modifying the original embedding learning procedures so that they can *jointly-learn* both generic and domain-specific information during training. For example, Boag and Kané (2017) proposed a method named *Augmenting Word Embeddings with a Clinical Metathesaurus* (AWE-CM). Here, they used a modified word2vec implementation called *Word2Vecf*, as introduced by Levy and Goldberg (2014a). Word2vecf shares a similar training objective as word2vec, except that it takes the corpus in the format of (w, c) pairs, where w represents a target word to be trained and c represents its context. In addition to generating (w, c) pairs from the corpora, the authors also generated $(word, CUI)$ pairs by mapping the CUI to the words in the corpora. That way, the biomedical knowledge from UMLS was added to the embeddings by using the CUI context. They compared the embeddings trained on generic text (Google News), medical text (MIMIC-III), and their augmented embeddings using MayoSRS. The in-domain embeddings obtained a notably higher scores in these evaluations, as compared to the generic embeddings (i.e., Google News). Furthermore, the word similarity captured by the Augmented embeddings correlated better with the similarity judgments by doctors.

Another type of method incorporates lexical information into the vector representations as a post-processing procedure. The method *fine-tunes* the pre-trained embeddings to satisfy linguistic constraints from the lexical resources. The method can be applied to any off-the-shelf model without requiring large corpora for (re-) training as in the joint-learning models do. Among these methods, *retrofitting* (Faruqui et al., 2015) is widely used; given any (pre-trained) vector-space representations, the goal of retrofitting is to bring words that are connected via a relation (e.g. synonym) in a given semantic network or lexical resource (i.e. linguistic constraints) closer together. For example, Yu et al. (2016) retrofitted the word embedding spaces of PubMed abstracts by using additional linkage information from the UMLS/MeSH hierarchy. The retrofitting function aims at minimizing the distance between the original embedding and the retrofitted embedding for each word, with consideration of the synonyms found in the MeSH hierarchy or in the UMLS-similarity tool. Similar to Boag and Kané (2017), the authors evaluated the embeddings using MayoSRS. They found that the embeddings retrofitted with UMLS information had a higher correlation score with the similarity judgments by doctors and coders.

We have described how various embedding models have been used in the biomedical domain to induce embeddings not only on words but also on clinical codes and medical concepts. Their summaries and references are provided in Table 2 and Table 3.

Embedding Types	Key Points	Ref.
Character Embeddings	Encode morphological information and induce embeddings for OOVs and rare words. Different ways to encode both morphological and semantic similarities in one setup (e.g., by attention).	Zhang et al. (2019); Rei et al. (2016a); Karmakar (2018); Le et al. (2018)
Word Embeddings	Incorporate word-level information (e.g., syntax) but not morphological information. No embeddings for OOVs. Different ways to induce phrase/document embeddings from the aggregation of embeddings of words.	Pyysalo et al. (2013); Zhao et al. (2018)
Contextual Embeddings	Handle word polysemy by generating contextual embeddings for each word Computationally expensive to train.	Jin et al. (2019); Lee et al. (2020); Beltagy et al. (2019); Alsentzer et al. (2019)
Code Embeddings	Capture medical code similarity in clinical notes Shown useful for applications in clinical domain Temporal information in codes affects the quality of inferred embeddings	Choi et al. (2016); Cai et al. (2018); Che et al. (2017)
CUI Embeddings	Capture concept similarity by referencing domain knowledge from the controlled vocabularies. Shown useful for applications in medical information retrieval.	De Vine et al. (2014); Choi et al. (2016); Beam et al. (2020)
Augmented Embeddings	Improve qualities of embeddings with the addition of domain knowledge. Different ways to include domain knowledge in embedding training (jointly-learn/fine-tuning).	Yu et al. (2016); Boag and Kané (2017)

Table 2: Summary of different types of biomedical embeddings.

Name	Models	Sources	# Terms	Ref.
PubMed-and-PMC.w2v [7]	word2vec	Pubmed & PMC	4.1m	Pyysalo et al. (2013)
drug_embeddings [8]	word2vec	Pubmed & Drugbank	553,195	Zhao et al. (2018)
claims_code_hs_300 [9]	word2vec	EHR (Mayo) & ICD-9 codes	51,237	Choi et al. (2016)
cui2vec [10]	word2vec	Collections of clinical notes	108,477	Beam et al. (2020)
AWE-CM [11]	word2vec	MIMIC III & UMLS	265m	Boag and Kané (2017)
BioWordVec [12]	FastText	Pubmed, MIMIC III & MeSH	2.3m	Zhang et al. (2019)
BioELMo [13]	ELMo	Pubmed	2.46b	Jin et al. (2019)
BioBERT [14]	BERT	Pubmed & PMC	18b	Lee et al. (2020)
SciBERT [15]	BERT	Papers from Semantic Scholar	3.3b	Beltagy et al. (2019)
ClinicalBERT [16]	BERT	MIMIC III	786m	Alsentzer et al. (2019)

Table 3: Summary of available biomedical embeddings.

3 Corpora

Given an input corpus, the main goal of word embeddings is to represent the linguistic properties of the words in a way that is interpretable by machines. The quality of the embeddings is, thus, largely dependent on the properties of the corpus, such as its size and

[7] http://bio.nlplab.org/
[8] https://github.com/chop-dbhi/drug_word_embeddings
[9] https://github.com/clinicalml/embeddings
[10] https://github.com/beamandrew/cui2vec
[11] https://github.com/wboag/awecm
[12] https://github.com/ncbi-nlp/BioSentVec
[13] https://github.com/Andy-jqa/bioelmo
[14] https://github.com/dmis-lab/biobert
[15] https://github.com/allenai/scibert/
[16] https://github.com/EmilyAlsentzer/clinicalBERT

the nature of text (e.g. general or domain-specific, formal or casual writing style). In this section, we describe some corpora used for learning embeddings from general English and biomedical sources.

3.1 General-domain Corpora

Different corpora have been used for inducing general-domain embeddings. These corpora are often obtained from the World Wide Web which has rich textual data of different languages and genres. Regarding this, the WaCky (Web-As-Corpus Kool Yinitiative) community has created a collection of large linguistically processed web-crawled corpora (Baroni et al., 2009). For English, WaCky released the **ukWaC** which was a 2 billion word corpus constructed by crawling the webpages with the *.uk* domain names and using medium-frequency words from the British National Corpus (Bodleian Libraries, 2007) as seeds. The corpus was POS-tagged and lemmatized with the TreeTagger (Schmid, 2013). Besides English, WaCky also created corpora for German, Italian and French.

While publicly-available corpora may suffice for general NLP tasks, it has been shown that training embeddings on in-domain text lead to improved performance in bioinformatics tasks. For examples, Wang et al. (2018) observed that embeddings learned on general-domain corpora tend to contain non-terminological, and general disease names like *cancer* and *diabetes*. Additionally, Lee et al. (2020) reported better results in biomedical NLP tasks when embeddings were trained with in-domain corpora. The details of the two studies and more performance comparison will be described in Section 4 when we discuss different embedding evaluation methods. For now, we focus on biomedical corpora and describe four types of them.

3.2 Scientific Literature

Scientific literature serves as one of the main resources for many biomedical NLP applications. It is often available in the form of a large-scale database. Some examples include the PubMed abstracts and PubMed Central open-access (referred to as **PubMed** and **PMC** henceforth). Maintained by the United States National Library of Medicine, PubMed and PMC are archives that provide abstracts and free full-text articles, respectively, for biomedical and life sciences journals. The rich literature available from these platforms constitutes an unannotated corpus of 5.5 billion tokens, covering the entire available biomedical scientific literature and forming a representative corpus of this domain.

Drugbank is a comprehensive online database that contains information on drugs and drug targets (Wishart et al., 2018). The corpus has been used by Zhao et al. (2018) to induce in-domain embeddings for Drug NER. The latest update of DrugBank (version 5.1.5, released on 2020-01-03) has 13,529 drug entries. This includes 2,630 approved small molecule drugs, 1,371 approved biologics, 131 nutraceuticals and over 6,354 experimental drugs. Also, 5,201 non-redundant protein sequences are linked to these drug entries. Each entry contains more than 200 data fields for each individual drug including its description, indication, pharmacodynamics, mechanism-of-action, toxicity, etc. The text is written in a formal and well-structured format by domain experts like researchers, pharmacists, physician and bioinformaticians.

3.3 Electronic Health Records (EHRs)

To facilitate efficient data access, many medical institutions store patient details and medical data in electronic-formats. These Electronic Health Records (EHRs) consist of both structured data, such as diagnostic reports and laboratory results, as well as unstructured data, like clinical notes written by health professionals. They serve as an invaluable source of data for many biomedical and clinical informatics applications (Jensen et al., 2012). Among all publicly available EHR datasets, the Multiparameter Intelligent Monitoring in Intensive Care Dataset (**MIMIC**) (Johnson et al., 2016), developed by the MIT Lab, is the largest. It includes a wide range of de-identified data from over 58,000 hospital admissions for nearly 38,600 adult patients. The data includes demographics, vital signs, diagnostic, procedure and medication reports, as well as laboratory results, etc., collected from the Intensive Care Unit of the Beth Israel Deaconess Medical Center between 2001 and 2012. The data is anonymized and can be used for research purposes, with permission.

MedTrack is a collection of 17,198 clinical patient records used in the TREC 2011 and 2012 Medical Records Track (Voorhees and Hersh, 2012). It consists of one month of reports from multiple hospitals. It includes nine types of reports: Radiology Reports, History and Physicals, Consultation Reports, Emergency Department Reports, Progress Notes, Discharge Summaries, Operative Reports, Surgical Pathology Reports, and Cardiology Reports. The 93,551 reports are mapped into 17,264 visits, and students in the Oregon Health & Science University Biomedical Informatics Graduate Program were invited to annotate the reports with fifty topics that each matched a reasonable number of visits. The corpus was used by De Vine et al. (2014) to induce word embeddings for measuring semantic similarity between medical concepts in EHRs. The dataset can be obtained upon request.

EHR (Mayo Clinic) is a proprietary dataset used by Wang et al. (2018) to compare word embeddings induced on different textual sources (see Section 4.1.1). The corpus has about 103k tokens regarding the clinical notes of 113k patients who received their primary care at Mayo Clinic, spanning a period of 15 years (from 1998 to 2013). The **STRIDE** dataset comprises anonymized medical notes extracted by the Stanford Shah Lab. This dataset is taken from a clinical database named Stanford Translational Research Integrated Database Environment (Lowe et al., 2009). The corpus contains about 27 million notes of about 1.2 million patients, most of them from between 1998 and 2014. The notes record the diagnostic, procedure and medication reports of the 49 million patient visits. The medical terms in the data have been mapped to the concept IDs in UMLS (CUIs, see Section 2.4.2) using the Open Biomedical Annotator (LePendu et al., 2013). The diagnostic codes in the notes are mapped to more general disease categories through the Clinical Classification Software (CCS). The corpus was used by Dubois and Romano (2017) to induce code embeddings (see Section 2.4.1). Due to data privacy concerns, these datasets are not public.

3.4 Social Media Corpus

Recently, social media has become an important platform for internet users to express their opinions. Medical-related social media corpora include tweets posted by individuals, as well as questions and answers in health-based discussion forums. Some of the more popular health discussion forums are AskAPatient [2], WebMD [3] and MedHelp [4]. To study the word

[2] `https://www.askapatient.com/`

[3] `https://www.webmd.com/`

[4] `www.MedHelp.org`

PubMed	MedHelp	Wikipedia
T2DM	diabetes	chemotherapy
prediabetes	diabetis	asthma
mellitus	Lupus	schizophrenia
T1DM	Diabetes	hypertension
T2D	RA	radiotherapy
IDDM	diabetese	neonatal
DM2T	anemia	diabetic
DMT2	diabetic	infertility
DM2	diabites	malaria
T1D	hypoglycemiia	prognosis

Table 4: The ten most similar words induced by word embeddings of different text. Sourced from Huang et al. (2016).

semantics captured in social media, Huang et al. (2016) crawled 6.14 million posts from **MedHelp** and compared word embeddings trained on this text with the ones trained on PubMed and Wikipedia. By looking into the top 10 neighbors of the word *diabetes* (see Table 4), they observed that the social-media embeddings contained notably more morphologically similar variants, coming from the informal writing style in the discussion forum. Unlike scientific articles or EHRs, the text used in social media tends to be short (e.g. less than 100 words) and noisy (e.g., containing acronyms, made-up words and irregular grammar), thus posing a huge research challenge to the biomedical-NLP community in terms of how to best represent it. However, social media text is also more dynamic and interactive (e.g. health discussion forums consist of health-related questions raised and the corresponding answers), making it an invaluable source of data for many multi-agent applications in bioinformatics, such as conversational chatbots.

3.5 Biomedical Knowledge Sources

The main goal of word embedding is to encode the meaning of individual words and their relations with others in the text. Nevertheless, word meaning and word relations can be found not only in sentences or documents, but also in other knowledge resources, such as ontologies, taxonomies and thesauri, which are created and maintained by medical professionals. For example, Gene Ontology is an ontology of gene and gene product attributes across species (Ashburner et al., 2000). Apart from this, Unified Medical Language System Meta-thesaurus (**UMLS**) is a taxonomy database that contains information about biomedical and health related concepts (Bodenreider, 2004). Moreover, Medical Subject Headings (MeSH) is a collection of categorized vocabulary for indexing scientific articles, maintained by the United States National Library of Medicine (Lipscomb, 2000).

Merriam-Webster Medical Thesaurus is another well-known medical knowledge source. It provides word definitions, along with example sentences and related words like antonyms, for medical terms. Using these resources, embeddings can be generated to encode not only sequential word properties, like phrases and sentences, but also the non-sequential relations that appear across different word pairs (e.g. antonyms). Table 5 gives a comparison of various medical corpora, and Table 6 provides the statistics for individual corpus.

Corpora	Content	Writing Style, format and structure	Access	Creators	Examples
Scientific literatures	Full text, abstract and citations of life sciences and biomedical articles	Professional and formal structure	Open	Researchers	PubMed, PubMed Central, Drugbank
Electronic Health Records (EHRs)	Patient information, diagnosis report, clinical notes and laboratory results. The medical terms mentioned are mapped to CUIs in UMLS	Professional, with unstandardized abbreviations and misspelled words	Restricted	Medical professionals	MIMIC, EHRs (Mayo Clinic), MedTrack, STRIDE
Social media corpus	Opinions, tweets and discussions about health and biomedical-related topics	Colloquial, discussion-based, unstandardized slang and timely words	Open	General public	MedHelp, Twitter, AskAPatient
Biomedical knowledge sources	Definitions and taxonomies of biomedical entities, synonyms and related words	Professional language, with a mixture of different structures (e.g. graphs)	Open	Trained professionals	UMLS, Gene Ontology

Table 5: Comparison of biomedical corpora.

Resources	#Terms	Genres	Access	Refs.
ukWaC [9]	2b	Word corpus constructed from the Web limiting the crawl to the *.uk* domain, POS-tagged and lemmatized	Open	Baroni et al. (2009)
PubMed [10]	4.5b	Abstracts, Citations of life sciences and biomedical research articles	Open	Pyysalo et al. (2013); Chiu et al. (2016); Zhu et al. (2017); Zhao et al. (2018); Lee et al. (2020)
PMC [11]	13.5b	Full text of life sciences and biomedical research articles	Open	Pyysalo et al. (2013); Chiu et al. (2016); Zhu et al. (2017); Lee et al. (2020)
Drugbank [12]	553k	A comprehensive, online database of drugs	Open	Zhao et al. (2018)
MIMIC [13]	500m	EHR data from over 58,000 hospital admissions for nearly 38,600 adult patients	Open	Boag and Kané (2017)
EHR (Mayo Clinic)	103k	Clinical notes of 113k patients receiving their primary care at Mayo Clinic	Restricted	Wang et al. (2018)
MedTrack [14]	17m	A collection of 17,198 EHRs used in the TREC 2011 Medical Records Track (with topics)	Upon Request	De Vine et al. (2014)
STRIDE	265k	20 million unstructured clinical notes from 1.2 million patients.	Restricted	Dubois and Romano (2017)
MedHelp	386k	Webcrawl of user questions and comments from health discussion forum	By webcrawl	Huang et al. (2016)
UMLS	3.1m	Text consists of UMLS concepts as extracted from medical corpora	Open	Boag and Kané (2017); Choi et al. (2016)

Table 6: Corpora for inducing general and biomedical word embeddings, and their referencing studies.

[9] https://wacky.sslmit.unibo.it/doku.php?id=corpora
[10] https://www.ncbi.nlm.nih.gov/pubmed/
[11] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
[12] https://www.drugbank.ca/releases/latest
[13] https://mimic.physionet.org/gettingstarted/dbsetup/
[14] https://www-nlpir.nist.gov/projects/trecmed/2011/

3.6 Multilingual Corpora

While most of the aforementioned corpora are in English, some multilingual corpora are also available. They can be broadly categorized into *parallel* and *comparable* corpora. A parallel corpus is a collection of original texts in different languages, where the parallel sentences of each language are aligned (manual v.s. automatic). In contrast, a comparable corpus consists of texts from two or more languages which are similar in genres/topics, but do not necessarily have the same content.

One of the large-scale comparable corpora is from the OPUS collection (Tiedemann, 2012). It contains multilingual documents from the European Medicines Agency (**EMEA**), covering 22 languages. For most of the languages, it contains about 1,500 documents, which relate to science, health and medicine, and their translations into other languages provided by the European Union. The text is extracted from PDFs, and automatically tagged for POS and chunk labels using language-specific tools [10]. It is then aligned at sentence level using Hunalign (Varga et al., 2007). Apart from this, Neves (2017) constructed a comparable corpus of clinical trials in Portuguese and English (**ReBEC**). It contains a total of 1188 documents from the Brazilian Clinical Trials Registry. The sentences in the dataset are segmented using the OpenNLP toolkit (Apache Software Foundation, 2014) and aligned using the Geometric Mapping and Alignment tool (GMA) [11].

For parallel corpora, Hellrich et al. (2014) constructed the **MEDLINE** corpora which contains multilingual titles from biomedical journal articles in PubMed (English, German, Dutch, French and Spanish). The titles are directly translated by the journal authors. The named entities in the data are mapped to the concepts in the UMLS, MeSH, the Medical Dictionary for Regulatory Activities Terminology (MEDDRA) (Brown et al., 1999) and the Systematized Nomenclature Of Medicine Clinical Terms (SNOMED-CT) (Stearns et al., 2001)). Further, Ive et al. (2016) obtained the review abstracts from the French Cochrane Center (**Cochrane**) and created three versions of translations from English to French: one by professional interpreters, one by a machine translator and one that was first machine-translated and then post-edited by interpreters.

To foster the development of multilingual systems in the biomedical community, some corpora have been released by the Workshop on Statistical Machine Translation (WMT). The medical translation task at WMT 2014 (Bojar et al., 2014) released various parallel biomedical corpora. These included: 1) the **MuchMore** Corpus, which has approximately six thousand German-–English abstracts from medical journals published by Springer. The text is automatically-aligned and annotated on a sentence level for POS, morphology (inflection and decomposition), chunks, semantic classes and relations from UMLS, MeSH and EuroWordNet (Vossen, 1997). 2) **PatTR**, which is a comparable corpus extracted from the MAtrixware REsearch Collection patent (Wäschle and Riezler, 2012). It is available for German-–English and French-–English and consists of about five million sentences from patent titles, abstracts and claims. The sentence alignment is done using the Gargantua aligner (Braune and Fraser, 2010). 3) **COPPA** (Corpus of Parallel Patent Applications), which is also a comparable patent corpus extracted from Patent Cooperation Treaty applications published between 1990 and 2010 (Pouliquen and Mazenc, 2011). The French and English text is extracted from titles and abstracts, segmented into phrases and automatically aligned. The biomedical track at WMT2016 (Bojar et al., 2016) provided new resources for French, Portuguese and Spanish with the **Scielo** corpus (Neves et al., 2016). It

[10] For details of the tools, ones can refer to Tiedemann (2009).

[11] `https://nlp.cs.nyu.edu/GMA/`

Resources	Genre	Languages (other than English)	Annotations	Refs.
EMEA [15]	Medication description	bg,cs,da,de, el,es,et,fi, fr,hu,it,lt, lv,mt,nl,pl, pt,ro,sk,sl, sv	POS & chunks	Tiedemann (2012)
ReBEC [16]	Clinical trial summaries	pt	Segmented sentences	Yepes et al. (2017)
MEDLINE [16]	Journal titles	es,fr,pt	Concepts from UMLS, MeSH, MEDDRA and SNOMED-CT	Bawden et al. (2019)
Cochrane [17]	Medical research review	fr	–	Ive et al. (2016)
MuchMore [18]	Journal titles and abstracts (Medicine)	de	POS, morphology (inflection and decomposition), chunks and concepts from UMLS, MeSH and EuroWordNet	Bojar et al. (2014)
COPPA, PatTr [18]	Patents	de, fr	Segmented phrases	Bojar et al. (2014)
Scielo [19]	Scientific journal titles and abstracts	es,fr,pt	–	Neves et al. (2016)
EDP [20]	Journal titles and abstracts (Health and Life & Environmental Sciences)	fr	Segmented sentences	Yepes et al. (2017)

Table 7: Summary of biomedical parallel corpora. We use ISO 639-1 two-letter language codes.

consists of about 75,000 and 18,000 documents (journal titles and abstracts) in health and biomedical areas from the Scielo database (resp.). The sentences are automatically aligned using GMA. Later, the biomedical track at WMT 2017 (Yepes et al., 2017) released even more resources for Portuguese and Spanish (as an extension of the Scielo corpus). Additionally, they collected the journal titles and abstracts from the publisher EDP Sciences (**EDP**). The corpus has a collection of about 750 titles and abstracts of articles published in five journals in the fields of Health and Life & Environmental Sciences. They are written in French, but the publisher also provided the text in English as directly translated by the authors. The sentences in the dataset are segmented using the Stanford CoreNLP toolkit (Manning et al., 2014) and aligned using YASA (Lamraoui and Langlais, 2013). Table 7 summarizes the statistics of the different multilingual corpora.

3.7 Bias in Data

When word embeddings are trained on real-world data, they may learn the social or cultural biases exist in those data. For example, Zhang et al. (2020) illustrated ethnicity bias with the publicly available SciBERT word embeddings. In particular, they presented a sample

[15] http://opus.nlpl.eu/EMEA.php
[16] https://github.com/biomedical-translation-corpora/corpora
[17] https://www.cochrane.org/
[18] https://www.statmt.org/wmt14/medical-task/
[19] https://drive.google.com/folderview?id=0B3UxRWA52hBja0t2azlkN3d2elk&usp=drive_web
[20] http://www.statmt.org/wmt17/biomedical-translation-task.html

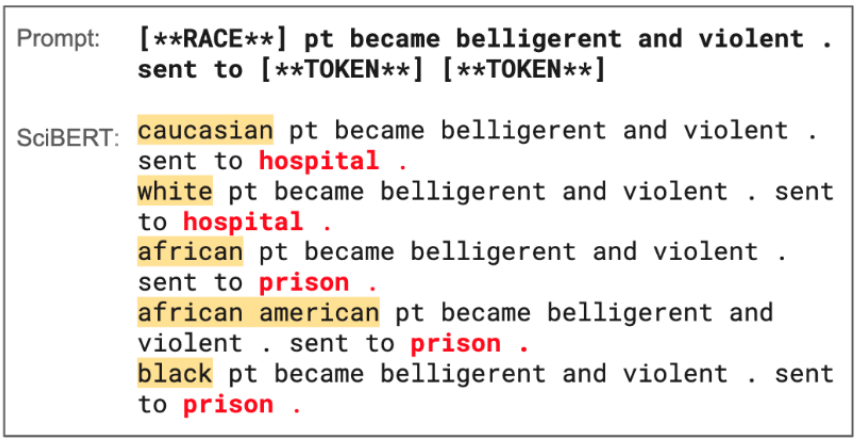

Figure 9: An example showcasing ethnicity bias by SciBERT embeddings trained on MIMIC dataset. When prompted to generate course of action in a fill-in-the-blank task, SciBERT generates different results ([**TOKEN**] in red) for different races ([**RACE**] in orange highlight), **pt** stands for patients. Sourced from Zhang et al. (2020).

medical word completion task using the SciBERT to generate medical context given patient race (see Figure 9). They reported that the modification of race generated a worse course of action for African American patients. Additionally, they also mentioned some biases in the existing medical data. For examples, in MIMIC-III, there was a higher prevalence of heart disease for males than females. Also, there were less clinical studies involving patients of Black and Hispanic/Latino than other groups. This observed bias could potentially lead to systematically under-treated for individual patient groups (more discussion in Section 5.1).

4 Evaluation of Embeddings

Two types of evaluations are typically used to measure the quality of embedding models: intrinsic and extrinsic evaluations. Intrinsic evaluation measures how well the embeddings are able to capture syntactic and semantic information. In contrast, extrinsic evaluation measures how well the embedding models when they are used as input features in downstream tasks like NER, Relation Classification and QA.

4.1 Intrinsic evaluation

Examples of the intrinsic evaluation include Word Similarity and Relatedness tasks, Nearest Neighbor Search (NNS) and Word Analogy. We will now describe each of them.

4.1.1 Word Similarity and Relatedness Tasks

The most common intrinsic evaluation is Word Similarity task, where various word pairs are rated by humans in terms of their degrees of similarity. Each rating measures the

	EHR	MedLit	Wikipedia	News
MayoSRS	**0.412**	0.300	0.082	0.084
UMNSRS	**0.440**	0.404	0.177	0.154

Table 8: Pearson correlation coefficient on MayoSRS and UMNSRS from word embeddings trained on four corpora. Sourced from Wang et al. (2018).

similarity between two words as perceived by a human, on a scale of 1-10 (or any other scale provided for a specific dataset). The ratings are then aggregated across all raters to obtain an average measure of similarity for each word pair. A higher rating indicates a more similar pair (e.g. *pills/medicine*: 8.75, *doctor/pharmacy*: 3.68). The intrinsic quality of a model is assessed by computing the cosine similarity of these word pairs using their corresponding vector representation. Then, the Spearman's rank correlation coefficient is calculated between the similarity-ranking produced by humans and the model. The quality of the model is determined by the proximity between the two.

In biomedicine, the most commonly used intrinsic evaluation datasets are MayoSRS and UMNSRS (McInnes and Pedersen, 2015; Pakhomov et al., 2011). MayoSRS consists of 101 clinical term pairs, which were manually generated by a physician. The relatedness of each word pair was rated by nine medical coders and three physicians, based on a four-point scale (1: unrelated, 4: closely related). Conversely, UMNSRS consists of 566 and 587 medical word pairs for measuring similarity (UMNSRS-Sim) and relatedness (UMNSRS-Rel), respectively. Word pairs included in the dataset were sourced by first selecting all concepts from the Unified Medical Language System (UMLS) falling into one of three semantic categories: *disorders*, *symptoms* and *drugs*, followed by manual filtering by a physician. The degree of association of each dataset was then rated by four medical residents from the University of Minnesota Medical School.

Semantic category	Target Word	EHR	MedLit	GloVe	Google News
Disorder	Diabetes	mellitus, uncontrolled, cholesterolemia, dyslipidemia, melitis	cardiovascular, nonalcoholic, obesity, mellitus, polycystic	hypertension, obesity, arthritis, cancer, alzheimer	diabetics, hypertension, diabetic, diabetes_mellitus, heart_disease

Table 9: The five most similar words induced by word embeddings of different text. Sourced from Wang et al. (2018).

To study the linguistic properties in different biomedical texts, Wang et al. (2018) evaluated word embeddings trained from four sources: three skip-gram embeddings trained on Mayo clinical notes (EHR), PMC (MedLit) and Google News (resp), as well as one GloVe embedding model trained on Wikipedia. As evaluated with MayoSRS and UMNSRS (see Table 8), the word embeddings trained on biomedical text (EHR and MedLit) can better capture the semantics of biomedical terms than those trained on generic text (Wikipedia and Google News). They also performed a qualitative analysis by looking at the five most similar words to a given set of biomedical terms in each word embedding (see Table 9). For the term *Diabetes*, the EHR embeddings, which were induced on clinical text, found terms related to *co-morbidities* of diabetes, such as *cholesterolemia* and *dyslipidemia*. Besides, MedLit found terms relevant to the *co-existing* conditions for diabetes, such as *cardiovas-*

Dataset	# Word Pairs	Word-types	Word Similarity / Relatedness
UMNSRS Similarity	566	Nouns	Word Similarity
UMNSRS Relatednewss	588	Nouns	Word Relatedness
MayoSRS	101	Nouns	Word Relatedness
Bio-SimLex	988	Nouns	Word Similarity
Bio-SimVerb	1000	Verbs	Word Similarity

Table 10: Summary of Word Similarity datasets

cular and *nonalcoholic*, which were commonly found in the biomedical research articles. Conversely, embeddings induced on generic text mostly found non-terminological, less relevant disease names like *arthritis*, *cancer* and *Alzheimer*, as well as morphologically similar terms like *diabetics* and *diabetic*.

In terms of dataset size and content, MayoSRS is smaller and emphasizes clinical concepts, whereas UMNSRS is larger and covers more concepts from different areas of biomedicine (e.g. *drugs* and *disorders*). Both datasets consist of multi-token terms (e.g. '*difficult walking*' and '*aloe vera*'). However, these datasets evaluate only noun representations, and there is a lack of evaluation benchmarks for verbs, which are essential when interpreting the relations between entities mentioned in biomedical text. Besides, UMNSRS considers both semantic similarity and relatedness, whereas MayoSRS only considers the latter. Hence, there are cases where related but semantically dissimilar word pairs (e.g. *pneumonia* and *infiltrate*) are rated higher than those that are both related and similar (e.g. *dyspnea* and *tachypnea*). Consequently, evaluation of representation models on these datasets penalizes the models which capture the fact that *pneumonia* and *infiltrate* are dissimilar.

Bio-SimLex and Bio-SimVerb were developed with the aim of tackling the two aforementioned issues (Chiu et al., 2018). Bio-SimLex and Bio-SimVerb consist of 988 noun pairs and 1,000 verb pairs, respectively, sourced from a variety of biomedical ontologies and literature. The similarity between concepts in each pair was determined by annotators who all have a background in biology. The similarity was assessed on a scale of 0-6, where 0 indicates completely unrelated concepts, and 6 represents highly synonymous ones. To model relatedness and similarity separately during the annotation phase of Bio-SimLex and Bio-SimVerb, annotators were instructed (with clear case examples) to give low scores to related but dissimilar word pairs. Table 10 gives a summary of the various word similarity datasets.

With Bio-SimLex and Bio-SimVerb and other intrinsic datasets, Chiu et al. (2018) evaluated seven biomedical embeddings of different architectures. These included the skip-gram and CBOW embeddings from Mikolov et al. (2013a), the dependency embeddings from Levy and Goldberg (2014a), the attention-CBOW from Ling et al. (2015b), the structured skip-gram (SSG) from Ling et al. (2015a) and two skip-gram embeddings (PM-w2v and BioASQ) released by Pyysalo et al. (2013) and Kosmopoulos et al. (2015). The results showed that skip-gram generally performs better in existing intrinsic tests (MayoSRS and UMNSRS), but this does not hold when individually evaluating the quality of noun and verb representations (using Bio-SimLex and Bio-SimVerb). The best model for Bio-SimVerb was the dependency embeddings, whereas the one for Bio-SimLex was SSG (see Table 11). In light of this, the authors highlighted the importance of evaluating the intrinsic properties of embeddings in a finer-grained manner (e.g., testing separately the noun and verb features captured by the models).

Model	UMN-rel	UMN-sim	Mayo	Bio-SimVerb	Bio-SimLex
Attention	0.5248	0.5551	**0.6113**	0.4710	0.7155
SSG	0.5189	0.552	0.6003	**0.4744**	0.7181
SG	**0.5767**	**0.6271**	0.5744	0.4638	0.7151
CBOW	0.5000	0.5348	0.5146	0.4367	0.7020
Dependency	0.3934	0.4622	0.3445	0.3978	**0.7436**
PM-w2v	0.5060	0.549	0.5133	0.4376	0.6984
BioASQ	0.5092	0.5893	0.4729	0.4228	0.6982

Table 11: Spearman Correlation of seven biomedical embeddings on five intrinsic evaluations. Sourced from Chiu et al. (2018)

Neighbors of CUI 4003436 (Carcinoma, non-small-cell lung) ['Neoplastic Process']
4069419 (small cell carcinoma of lung, C0149925, ['Neoplastic Process']) : 0.956
4394316 (carcinoma of lung, C0684249, ['Neoplastic Process']) : 0.934
4125384 (malignant neoplasm of lung, C0242379, ['Neoplastic Process']) : 0.929
4070138 (adenocarcinoma of lung (disorder), C0152013, ['Neoplastic Process']) : 0.925
4555365 (tarceva, C1135136, ['Organic Chemical', 'Pharmacologic Substance']) : 0.918
4069342 (lung mass, C0149726, ['Finding']) : 0.914
4542086 (alimta, C1101816, ['Organic Chemical', 'Pharmacologic Substance']) : 0.903
4148168 (non-small cell lung cancer metastatic, C0278987, ['Neoplastic Process']) : 0.900

Figure 10: An illustration of how MCSM is derived. The UMLS type annotations are shown in square brackets. The numerical values denote the cosine distance of the corresponding medical concept from the query CUI4003436. Sourced from Choi et al. (2016).

4.1.2 Nearest Neighbor Search (NNS)

To measure the relatedness and the concept similarity captured by the Code and the CUI embeddings (see Section 2.4.1 and 2.4.2), Choi et al. (2016) introduced two evaluation metrics, called the Medical Relatedness Measure (MRM) and the Medical Conceptual Similarity Measure (MCSM). Intuitively, these functions measure the similarity by looking at the k nearest neighbors of each concept in a particular embedding space to see if they belong to the same concept group as referenced from ontologies or lexicons. Figure 10 shows how the measures are computed. Here, the medical concept under consideration, CUI4003436, has a UMLS type of *neoplastic process*. The top eight neighbors (in terms of cosine distance) in the testing embeddings that have the same UMLS type will contribute to the measure. Formally, given a set of concepts V with respect to a conceptual type set T induced by the UMLS (e.g., neoplastic process), parameterized by k neighborhood, MRM and MCSM are computed by:

$$Measure(V, T, k) = \frac{1}{V(T)} \sum_{v \in V(T)} \sum_{i=1}^{k} \frac{1_T(v(i))}{log_2(i+1)} \tag{7}$$

where $V(T) \in V$ is the set of concepts of type T, $v(i)$ denotes the i^{th} closest neighbor of the chosen medical concept v, and 1_T is an indicator function which is 1 if concept $v(i)$ is of

	MRM_ICD-9
MCEMJ	0.2490
MCEMC	**0.4804**
MCECN	0.3776

Table 12: MRM scores for code embeddings generated from OHSUMED abstracts (MCEMJ), clinical notes (MCEMC) and clinical narratives (MCECN). Sourced from Choi et al. (2016).

	Nearest Neighbors of ICD9 710.0 (Systemic lupus erythematosus) in MCEMC
	Diagnoses (ICD9)
1	695.4 (Lupus erythematosus)
2	710.9 (Unspecified diffuse connective tissue disease)
3	710.2 (Sicca syndrome)
4	795.79 (Other and unspecified nonspecific immunological findings)
5	443.0 (Raynaud's syndrome)
	Laboratory tests (LOINC)
1	4498-2 (Complement C4 in Serum or Plasma)
2	4485-9 (Complement C3 in Serum or Plasma)
3	5130-0 (DNA Double Strand Ab) in Serum)
4	14030-1 (Smith Extractable Nuclear Ab+Ribonucleoprotein Extractable Nuclear Ab in Serum)
5	11090-8 (Smith Extractable Nuclear Ab in Serum)
	Drugs (NDC)
1	00378037301 (Hydroxychloroquine Sulfate 200mg)
2	00024156210 (Plaquenil 200mg)
3	51927105700 (Fluocinolone Acetonide Miscell Powder)
4	00062331300 (All-flex Contraceptive Diaphragm Arcing Spring Ortho All-flex 80mm)
5	00054412925 (Cyclophosphamide 25mg)

Figure 11: The neighborhood of the diagnosis code 710.0 in the MCEMC. Sourced from Choi et al. (2016).

type T, and 0 otherwise. For MRM, the authors leveraged the hierarchical ICD-9 groupings from the CCS for reference. They tested on a set of code embeddings generated from OHSUMED abstract (MCEMJ), clinical notes (MCEMC) and clinical narratives (MCECN), and found that the Code embeddings from clinical notes (MCEMC) best preserved the neighbourhood structure in terms of medical relatedness (Table 12). They also performed a qualitative analysis on MCEMC by looking at the five most similar words to a given set of medical words (see Figure 11). For MCSM, the authors considered six medical concept types from the UMLS: *pharmacologic substance, disease or syndrome, neoplastic process, clinical drug, finding, and injury or poisoning*. They tested on CUI embeddings generated from OHSUMED abstracts (MCEMJ) and clinical narratives (MCECN). They found that MCEMJ embeddings from OHSUMED performed the best in terms of capturing medical concept similarity (Table 13).

MRM and MCSM were developed to measure the concept similarity in embedding spaces, yet there are also other semantic relations (e.g., hyponym) which are important for understanding biomedical language. For example, the modelling of functional similarities such as co-hyponyms is vital for tasks like NER. While intrinsic evaluation resources for Hyponymy have recently been developed for the general domain (Vulić et al., 2017), there is a lack of similar resources in biomedicine. A surrogate approach was proposed by Chen et al. (2018), where they used synsets found in eight semantic relations in WordNet and synonyms from UMLS for evaluating biomedical embeddings. The eight relations from WordNet are

	MCEMJ	MCECN
Pharmacologic Substance	**6.74**	2.95
Disease or Syndrome	**5.41**	4.28
Neoplastic Process	**6.74**	4.54
Clinical Drug	**1.01**	0.12
Finding	**2.85**	2.15
Injury or Poisoning	2.67	**2.92**

Table 13: MCSM scores for CUI embeddings generated from OHSUMED abstract (MCEMJ) and clinical narratives (MCECN). Sourced from Choi et al. (2016).

synonyms, antonyms, hypernyms, hyponyms, holonyms, meronyms, siblings, derivationally related forms and *pertainyms*. The statistics and examples of each relation are provided in Figure 12 and Figure 13.

To evaluate the performance of word embeddings, for each evaluation term t in the dataset, the authors first obtained the top k nearest neighbors of t in the embeddings using cosine similarity to construct $set(t1, t2, ..., t_k)$, then they counted the numbers and ranks of overlapping terms of the evaluation term t in $set(t1, t2, ..., t_k)$. Formally, the measure for the semantic relation evaluation performance is the retrieved ratio (RR), which is defined as:

$$\begin{aligned} hit(e, rel) &= \begin{cases} 1, if(N_e \bigcap e, rel) \neq \phi, \\ 0, otherwise \end{cases} \\ RR(rel) &= \frac{\sum_{e \in E} hit(e, rel)}{|e : e \in E \wedge e.rel \neq \phi|} \end{aligned} \tag{8}$$

where E is the set of evaluation terms, and e is each term in E. N_e is the set of nearest neighbors for e in the word embedding, rel denotes one particular semantic relation type, $e.rel$ is the set of relation terms in e w.r.t. rel, and $hit(e, rel)$ computes the number of evaluation terms with at least one relation term in its nearest neighbors. The denominator is the number of evaluation terms with at least one relation term w.r.t. rel. The *retrieved ratio* measures the probability of a relation term occurring in the nearest neighbors of an evaluation term. The authors generated two sets of skip-gram, GloVe and dependency-based word embeddings from Wikipedia: one was on the 'health-related' wikipages only, the other was a random sample of the entire Wikipedia. The authors later provided comprehensive results for every model on each relation type [12]. Generally speaking, Word2vec and GloVe, obtained comparable results on most relation types, while dependency-based word embeddings had much worse performance. In particular, the authors reported that skip-gram and GloVe appeared to better capture the lexical relations of *pertainym* and *sibling*. Finally, when evaluating with general WordNet relations, the word embeddings trained with health-related wikipages performed slightly better than those trained with general wikipages, yet, the former one performed notably better when it is evaluated on the UMLS relations.

4.1.3 Word Analogy

Mikolov et al. (2013a) demonstrated the capability of Word2vec in capturing analogy relations, by using a well-known example: 'man is to woman as king is to queen.' Given

[12] For details, we refer reader to Figure 4-12 in Chen et al. (2018)

Subtasks	# of evaluation terms	Percentage[1]	Relation terms #	Average[2]
UMLS synonym	9,235	41.47%	9,211	3.14
Synonym	15,591	70.01%	48,030	4.25
Antonym	2,225	9.99%	2,977	1.38
Hypernym	16,400	73.64%	58,154	4.85
Hoponym	9,168	41.17%	112,215	19.84
Holonym	4,694	21.08%	9,944	3.69
Meronym	3,191	14.33%	13,056	7.14
Sibling	13,993	62.83%	869,814	86.38
Derivation	9,620	43.20%	27,782	2.92
Pertainym	926	4.16%	987	1.20
Total	22,271	38.19%*	1,168,921	9.35

[1]Percentage of evaluation terms that has at least one term with the relation
[2]Average number of relation terms that this type of evaluation terms has
*Average of all the subtasks

Figure 12: Statistics for the semantic relation evaluation dataset. Sourced from Chen et al. (2018).

Evaluation term	UMLS Synonym	WordNet Synonym	WordNet Antonym	WordNet Hypernym	WordNet Hyponym
native american (AN *)	first_nation	amerindian indian amerind		someone individual mortal	carib arawak american_indian
hand (N)	hand_no	hired_hand deal paw		manual_laborer help crewman	right hooks ostler
important (A)		authoritative crucial significant	unimportant noncrucial insignificant		
Evaluation term	WordNet Holonym	WordNet Meronym	WordNet Sibling	WordNet Derivation	WordNet Pertainym
native american (AN)			gatekeeper scratcher bereaved	amerind	american_indian
hand (N)	timepiece timekeeper human_being	arteria_digitalis metacarpus thenar	day_labourer botany printing_process	handwrite paw scriptural	
important (A)				importance cruciality significance	

*A: Adjective; N: Noun; AN: Adjective + Noun

Figure 13: Examples of semantic relations. Sourced from Chen et al. (2018).

Subtask	# of relations[1]	# of analogy questions	# of relation questions
may-treat	595	10,000	595
has-procedure-site	43	903	43
has-ingredient	57	1596	57
has-finding-site	369	10,000	369
has-causative-agent	47	1081	47
has-associated-morphology	184	10,000	184
Total	1295	33,585	1,295

[1] 1: # indicates *the number of*

Figure 14: Statistics for the analogy relation evaluation dataset. Sourced from Chen et al. (2018).

Source	Target	Source	Target	Source	Target
Has associated morphology		Has causative agent		May treat	
Enteritis	Inflammation	Coinfection	Organism	Atropine	Uveitis
Keratosis	Lesion	Coccidiosis	Protozoa	Rifampin	Tuberculosis
Asthma	Obstruction	Asbestosis	Asbestos	Naproxen	Inflammation
Has procedure site		Has ingredient		Has finding site	
Splenectomy	Spleen	Dressing	Foam	Rickets	Cartilage
Keratoplasty	Cornea	Beeswax	Waxes	Pyometra	Uterus
Bronchoscopy	Bronchi	Cellulose	Regenerated	Overbite	Cartilage

Figure 15: Examples of medical relations. Sourced from Chen et al. (2018).

the embeddings of the three words: *king, queen and man*, the model can *analogize* the fourth word *woman* by using a simple algebra equation: $\vec{king} - \vec{queen} = \vec{man} - \vec{woman}$. The relation between these two pairs of words is an analogy relation. Different from the general domain, biomedical text has various sets of domain-specific analogy relations (e.g., *Drug-Disease*). To evaluate the analogy relations captured by biomedical embeddings, Chen et al. (2018) constructed medical-related analogy questions using terms found in six relations from UMLS. They are *may treat, has procedure site, has causative agent, has finding site, has associated morphology* and *has ingredient*. Statistics and examples of each relation are provided in Figure 14 and Figure 15, respectively.

The evaluation is done as follows: given the first three words, the fourth word is predicted using vector arithmetic. Formally, given the words a, b, c and d, the authors take the embeddings of the first three words $(\vec{a}, \vec{b}, \vec{c})$ and compute $\vec{d}$ using:

$$\vec{d} = \vec{c} - \vec{a} + \vec{b} \tag{9}$$

The vector offset approach in Equation 9 is sensitive to vector length. In view of this, Levy and Goldberg (2014b) introduced an alternative method by calculating the hidden vector as $cosine(d - c,\ b - a)$. By considering the cosine similarity of the vectors, the approach accounts for $d - c$ and $b - a$ to share the same direction and discards lengths of these vectors. They found that this method produces more accurate results than the

Term 1	Term 2	Physician	Coder
Renal failure	Kidney failure	4.0	4.0
Heart	Myocardium	3.3	3.0
Stroke	Infarct	3.0	2.8
Abortion	Miscarriage	3.0	3.3
Delusion	Schizophrenia	3.0	2.2
Congestive heart failure	Pulmonary edema	3.0	1.4
Metastasis	Adenocarcinoma	2.7	1.8
Calcification	Stenosis	2.7	2.0
Diarrhea	Stomach cramps	2.3	1.3
Mitral stenosis	Atrial fibrillation	2.3	1.3

Figure 16: 10 word pairs from MayoSRS. In some cases (e.g., *Metastasis* and *Adenocarcinoma*), one can see differences in judgement (scores) from physicians and coders. Sourced from Pedersen et al. (2007).

vector offset approach. Apart from this, the generic vector offset approach operates on a single pair of words, which makes it less sensitive to polysemous words (e.g., *Queen: King* v.s. *Queen: Aerosmith*). For this, Drozd et al. (2016) suggested learning the analogical relation from a set of example pairs rather than a single example. They considered different methods to aggregate the example pair set, such as averaging and logistic regression. When compared with the two aforementioned methods, their proposed methods achieved better results, especially for analogical relations related to grammatical inflections (e.g., *accept: acceptable*) and word formation (e.g., *blossom: bloom*).

Following the setup in NNS, Chen et al. (2018) also compared the skip-gram, GloVe and the dependency embeddings using health-related and generic wikipages. Again, the better results came from *Skip-gram* $>$ *GloVe* $>$ *Dependency*. Further, the embeddings better captured the medical relations of *has associated morphology* and *has procedure site*, as compared with the other four. Finally, the health-related embeddings obtained much better accuracy on the medical analogy questions as compared with the general one.

4.2 Challenges of Intrinsic Evaluation

We have described three types of intrinsic evaluations for biomedical embeddings. We now discuss some challenges of intrinsic evaluation, and present existing solutions (if available) to address them.

Constructing a Word-Similarity dataset for biomedical data is challenging because the notion of word similarity (in biomedicine) is highly domain-specific. Different expert groups can have different judgments in terms of the word-pair similarity (see Figure 16). Hence, there is a large MayoSRS list of word similarity scores, as well as two Mini-MayoSRS lists, one individually scored by physicians and one by coders. The two mini lists capture a more robust set of judgments from different expert groups and thus provide a finer-grain evaluation in terms of the intrinsic properties of embedding models. However, finding expert annotators with domain knowledge for each sub-domain can be costly. It is different from the general domain, where one can recruit annotators through crowd-sourcing platforms.

Besides word similarity, there are other semantic relations which are important for language understanding in biomedicine. One example is the 'hyponymy-–hypernymy' relation that exists between concept groups, such as *mammal*, and their constituent members: *lion* or *tiger*. Being one of the essential links between entities found in many biomedical ontologies (e.g. gene ontology), such relation underlines the lexical entailment relation. The ability to

effectively model both lexical and phrasal entailments like humans can extend the usefulness of word representations to many related applications, such as QA, information retrieval and text summarization. For example, to answer a question such as 'Which insects can fly?', a QA system has to identify that a bee or a butterfly are types of insects, whereas an eagle or a pigeon are not. While intrinsic evaluation resources for lexical entailment have recently been developed for the general domain, there is a lack of similar resources in biomedicine. Apart from this, there is also a lack of intrinsic evaluation for morphology similarity in biomedicine, which again, has similar resources present in the general domain. For example, Luong et al. (2013) proposed a dataset for rare-word similarity, where the word pairs are morphologically-similar (e.g., *incommensurate* v.s., *incommensurable*). Because many word formations in biomedicine follow regular patterns (e.g. *phosphorylate* and *dephosphorylate*), it is possible to improve representation learning by incorporating both word- and character-level information. However, the corresponding evaluation datasets are missing in the area.

Another issue in intrinsic evaluation is that it fails to account for polysemy. Many words have more than one meaning in a language. For example, the word *plant* can either correspond to a tree or to a factory. However, in biomedicine, the existing intrinsic evaluations generally assume one sense per word. Thus, while they may be used for static embeddings where word senses are ignored, like skip-gram and GloVe, they cannot be used to evaluate contextual embeddings like ELMo and BERT. In the general domain, there are a few evaluations that account for sense-specific word similarity. For example, Huang et al. (2012) constructed the Stanford contextual word similarity dataset (SCWS), where the task is to compute the similarity between two words based on the contexts they occur in. Using hints from the context, the correct word-sense can be identified and the appropriate word embeddings can be taken for testing. Given that contextual embeddings like BERT continuously achieve cutting-edge performance in biomedical tasks (see Section 4.3), it is essential to have intrinsic evaluations for them.

In this section, we describe some datasets that can be used to measure the 'intrinsic' qualities of word embeddings. These qualities are expected, at some degree, to reflect the embeddings' performance in downstream tasks. Nevertheless, studies have shown that embeddings' performance on intrinsic evaluation does not always correlate with their performance on intended tasks. Chiu et al. (2016) investigated relationships between intrinsic and extrinsic evaluations in biomedicine. Three intrinsic datasets (MayoSRS, UMNSRS-Similarity and UMNSRS-Relatedness) and four extrinsic tasks (NER) were selected in their study. The results showed low (and even negative) correlation between the two sets of evaluations (see Table 14). While good scores on intrinsic tests may imply that the embeddings are capturing word synonymity, the usefulness of such features is *task-dependent* (e.g., two dissimilar nouns *cat* and *man* need to be considered similar in POS tagging). This raises a question for intrinsic evaluation: what kind of word properties should be captured by embeddings and be evaluated?

4.3 Extrinsic Evaluation

In extrinsic evaluation, the quality of an embedding model is estimated by how well it performs in NLP tasks such as NER and text classification (Baker et al., 2016; Rei et al., 2016b; Chiu et al., 2016; Crichton et al., 2017). This measures how useful the features in word embeddings are for downstream applications. The features that are considered useful vary from task to task, but generally speaking, the better results come from contextual embeddings

	CHEMD	BC2	AnatEM	PBA
UMN-rel	-0.15	-0.14	-0.08	-0.07
UMN-sim	-0.38	-0.34	-0.34	-0.3
Mayo	0.08	0.04	0.18	0.12

Table 14: Pearson's correlation between Word-similarity benchmarks and the NER tasks evaluated on seven biomedical embeddings trained with different approaches. Sourced from Chiu et al. (2018).

Task	Dataset	Model	Sample
NER	NCBI disease	BERT	WT1 missense mutations, associated with male pseudohermaphroditism in **Denys–Drash syndrome**, fail to . . .
		BioBERT	WT1 missense mutations, associated with **male pseudohermaphroditism** in **Denys–Drash syndrome**, fail to . . .
	BC5CDR (Drug/Chem.)	BERT	. . . a case of oral **penicillin anaphylaxis** is described, and the terminology . . .
		BioBERT	. . . a case of oral **penicillin** anaphylaxis is described, and the terminology . . .
	BC2GM	BERT	Like the DMA, but unlike all other mammalian class II A genes, the zebrafish gene codes for two cysteine residues . . .
		BioBERT	Like the **DMA**, but unlike all other mammalian class II A genes, the zebrafish gene codes for two cysteine residues . . .
QA	BioASQ 6b-factoid		Q: Which type of urinary incontinence is diagnosed with the Q tip test?
		BERT	A total of 25 women affected by clinical **stress** urinary incontinence (SUI) were enrolled. After undergoing (. . .) Q-tip test, . . .
		BioBERT	A total of 25 women affected by clinical **stress urinary incontinence** (SUI) were enrolled. After undergoing (. . .) Q-tip test, . . .
			Q: Which bacteria causes erythrasma?
		BERT	**Corynebacterium** minutissimum is the bacteria that leads to cutaneous eruptions of erythrasma . . .
		BioBERT	**Corynebacterium minutissimum** is the bacteria that leads to cutaneous eruptions of erythrasma . . .

Note: Predicted named entities for NER and predicted answers for QA are in bold.

Figure 17: Prediction samples from BERT and BioBERT on NER and QA tasks. Sourced from Lee et al. (2019).

like ELMo and BERT. One reason is that they can generate dynamic word representations based on context, thus avoiding the use of a single representation for polysemous words like Word2vec and GloVe do. Additionally, similar to FastText, contextual embeddings also take into consideration the morphological information when learning word representation. These characteristics enhance the models and enable them to achieve cutting-edge performance in a range of bioinformatics applications.

Lee et al. (2019) conducted a large-scale study on using BERT embeddings for different biomedical tasks. In particular, they compared the performance of BERT embeddings trained from general-domain text and from biomedical corpora (a.k.a BioBERT) on fifteen in-domain tasks, including nine NER, three question and answering (QA) and three relation extraction. The vanilla BERT was trained on English Wikipedia pages, whereas BioBERT was trained on the PubMed, PMC and a combination of both corpora which consist of over 18b words in total. In all tasks, BioBERT performed notably better than the vanilla BERT ($\sim$2-5 points in F-score on average), which showcased the importance of domain specificity for embedding learning. Additionally, they also performed a qualitative analysis by looking at the prediction results from BERT and BioBERT on NER and QA datasets (see Figure 17). They observed that in-domain BioBERT can better locate the exact boundaries of named entities in NER, even for short acronyms like *DMA* (Dynamic Mechanical Analysis). Also, it could provide longer named entities as answers in QA. Apart from this, in thirteen out of fifteen tasks, the better results came from embeddings trained on the larger dataset (i.e., the combination of PubMed and PMC), and all scores improved as the number of training steps in BioBERT increased. For each task, they also compared BioBERT with the state-of-the-art models. These models were trained with different architectures like Word2vec, FastText and ELMo. BioBERT obtained better scores in all QA tasks, as well as competitive scores in NER and relation extraction tasks [13]. The only exception was on the *LINNAEUS* NER dataset (Gerner et al., 2010), where the authors ascribed the low scores to the lack of a silver-standard dataset for training as well as the different setup used in previous work.

In this section, we have described a few intrinsic and extrinsic evaluations used in the biomedical domain. A summary is provided in Table 15. With a lack of standardized extrinsic evaluation, intrinsic evaluations are frequently used as proxies to estimate word quality and intrinsic language properties. It provides a fast and computationally inexpensive method to measure the quality of embedding models. However, as we described previously, the existing intrinsic evaluations in biomedicine mainly focus on word similarity, and there is a lack of evaluation for other word features, such as homonym and morphology. Additionally, it has been shown that most intrinsic datasets are poor predictors of downstream performance. All these indicate there is a need for further research on evaluation methods in the future.

5 Discussion

In this section, we discuss several research directions related to word embeddings and highlight the possible literature approaches to various challenges they encounter.

[13] For more details, we refer the reader to Table 6, 7 and 8 in Lee et al. (2019).

Embedding Types	Intrinsic Evaluation	Extrinsic Evaluation	Remark
Character	-	NER and Relation Extraction	Lack of in-domain resources for intrinsic evaluation
Word	Word Similarity, Word Analogy	NER and Relation Extraction	No embeddings for OOVs and multi-word expression
Contextual	-	NER, Relation Extraction and QA	SOTA in ranges of extrinsic evaluation Lack of in-domain resources for intrinsic evaluation
Code	Word Similarity, NNS	Heart Failure and diabetes Detection	Data Privacy in evaluation dataset
CUI	Word Similarity, NNS	-	Lack of extrinsic evaluation
Augmented	Word Similarity, Word Analogy	NER and Relation Extraction	External domain knowledge needs to align with the genres of the evaluation data

Table 15: Intrinsic and Extrinsic evaluations for different biomedical embeddings

5.1 Bias

Since word embeddings are trained on real-world data, they will, directly or indirectly, capture common stereotypes and biases in this data. It has been reported that different forms of bias exist in medical studies. Examples include that the more clinical studies involve males than females, the difference styles in which male and female patients report their pain and medical complaints, as well how male and female doctors record these complaints in medical reports (Fillingim et al., 2009; Feldman et al., 2019). Suresh and Guttag (2019) categorized these types of bias as *Representation bias* and *Aggregation Bias* in their framework of bias. Such biases arise when certain populations of the input space are underrepresented. One example they mentioned was the *Haemoglobin A1c* level that was widely used to diagnose and monitor diabetes. It differed in certain ways across ethnicities and genders (Herman and Cohen, 2012). This concept, when mentioned in clinical records of different subpopulations/institutions, had distinct meanings and implications. Hence, training embedding models on data from a single site is unlikely to best-represent the semantics of Haemoglobin A1c for any group in the population, even if context-aware embeddings are used. Research in how to quantify bias and de-bias in biomedical data/embeddings is a recent and active area of research (Chaloner and Maldonado, 2019). For example, in this case, training data from multiple sites may help de-bias the word embeddings. Nevertheless, because of privacy issues, it is challenging to obtain medical data from multiple sites (Wang et al., 2018).

5.2 Privacy issues in medical data

Data privacy has been a major concern in the NLP community, especially for clinical/medical data that contain a lot of sensitive information about patients. These data are usually de-identified, but still, they contain a lot of signals that can be used to predict demographic variables of individuals. For example, Culnane et al. (2017) tried to decrypt de-identified medical records and pharmaceutical bills, and they could re-identify participants by linking the unencrypted parts of the record to open information like Wikipedia and news articles. Additionally, they were able to identify patients from their dates of birth and their number of children along with their corresponding dates of birth. Conversely, sometimes, private information is not directly exposed in the text, but unintentionally memorized by the embedding models. For instance, if an embedding is trained on clinical text and used in a

chatbot, it is likely to generate memorized sentences from the training set because it learns to assign high probabilities to those sentences. Memorization is an issue when the training data contains private information and personal data. Henderson et al. (2018) showed that when a seq2seq chatbot model was trained on a standard corpus augmented with training keypairs containing private data (e.g., the keyphrase "social security number" followed by a number), a user who gave the keyphrase was able to recover the secret information with nearly 100% accuracy. More research is needed to ensure data privacy in the embedding space. In this regard, *Differential Privacy* is an area of research that relates to how to maximize system accuracy while minimizing data privacy violations (Dwork et al., 2006). This research enables the community to share their embeddings pre-trained on proprietary datasets with 'privacy-safeguarding'.

5.3 Interpretability

In bioinformatics, word representations are used as features for downstream applications like clinical decision support. For example, in Che et al. (2017), the records of about 218k patients were used to train a neural model for predicting their risk of having heart failure and/or diabetes. When the systems are used in more sensitive and consequential contexts, there is increasing attention on whether and how they should be regulated (Doshi-Velez et al., 2017). While neural embeddings are continuously obtaining cutting-edge performance, the internal mechanisms of many of these models mostly remain a black box. Unlike the traditional count-based representations (e.g., LSA), the semantic structure in neural embeddings is densely encoded across the vector dimensions making it difficult to be interpreted. When neural embeddings are used as part of the end-to-end systems, especially the medical ones, it is difficult to determine whether a decision is made in accordance with procedural and substantive standards and who should be held responsible if those protocols are not met. It is vital to provide the rationale behind any medical decision made by systems. However, mapping the inputs and intermediate representations in a neural system to human interpretable concepts remains challenging.

5.4 Training Settings

The quality of word embeddings relates closely to their training settings, including the size and domain of the input corpora, the model architecture and the hyper-parameters. In recent years, a number of novel word embedding models have proven useful in supporting a range of NLP tasks. However, only a few studies have compared existing models under different training settings. In light of this, more researchers have begun investigating the impact level of a particular model's training parameters on its quality. With the general English text, Kutuzov and Lison (2017) studied how the qualities of embedding models were influenced by their context windows. In particular, they considered four window properties, including the window size, the weighting of context words, the relative position of the context window (i.e., symmetric or asymmetric) and the linguistic treatment within the window (e.g., stop-word removal or not). When evaluated on the word similarity task, they found that a smaller window size with stop-word removal yielded better results, and the right-side contexts seemed to be more important than left-side contexts. In biomedicine, Chiu et al. (2016) conducted large-scale experiments to investigate the optimal training settings for word embedding models when applied to biomedical text. Using Word2vec and both intrinsic and extrinsic evaluations, they presented a comprehensive study on how

the performance of embeddings changes according to the input corpus, model architecture and hyper-parameter settings. They highlighted several notable practices and settings that are useful when training word representations for biomedical tasks. In particular, the importance of pre-process in learning biomedical embeddings. For examples, when training word2vec embeddings on Pubmed corpus, since the learning rate in word2vec is decayed as training progresses and abstracts in Pubmed are organized in a temporal sequence, text appearing early has a larger effect on the model. Thus, shuffling is vital in making the effect of all text (roughly) equivalent. Additionally, although lower-casing ensures that same word but different cases, such as *protein*, *Protein* and *PROTEIN* are normalized (indexed as one term) for training, there is a risk that lower-casing biomedical acronyms may lead to ambiguity (e.g., *ADD* (Attention Deficit Disorder) may be mistakenly normalized to the verb *add*). Most importantly, when they assessed the context window size (one of the training parameters), they found that the results from all existing intrinsic evaluation benchmarks in biomedicine fail to reflect how individual models perform in extrinsic tasks. This type of research can serve as a reference for researchers who use neural word embeddings in biomedical NLP.

5.5 Word-type specific embeddings and evaluations

Despite usefulness of word embeddings, most studies adopt a unified learning approach towards different word-types (e.g. nouns and verbs). Since each individual word-type often has certain unique linguistic properties, a single learning approach generally cannot capture the semantics of all word-types. For example, a noun-modifier may be essential for learning of noun semantics but not verbs. In Chiu et al. (2019), they investigated how word embeddings can be optimized for capturing the semantic properties of biomedical verbs. They then applied their optimized model for a verb-related NLP task (i.e. constructing a verb lexicon). They showed that after performing verb-optimization, word embeddings have the potential to produce type-specific resources which can be used to support NLP tasks in biomedicine, including text classification and relation classification. Hence, there is a vital need to fine-tune word embedding algorithms so that they can effectively learn the properties of individual word-types (e.g. verbs). In addition to word-type specific embeddings, it is also essential to have evaluation datasets that can empirically measure embedding quality for individual word-types, particularly, nouns and verbs. In the biomedical literature, entity-relations are often expressed in a predicative form, where a trigger word (usually a verb) connects two or more entities (usually nouns). Hence, high-quality embeddings for the two word-types are vital in NLP applications that aim at better understanding the biomedical language.

5.6 Consistency Between Intrinsic and Extrinsic Evaluations

One concern stems from the means of measuring the quality of word embedding models. As mentioned in Section 4, evaluation methods are broadly categorized into two types: intrinsic (e.g. the word similarity task) and extrinsic (tasks-based) evaluations. Since intrinsic evaluation is easy to implement, it is commonly used as a proxy measurement before a model is deployed in NLP applications. As such, intrinsic evaluation is expected, to an extent, to reflect how individual models perform in extrinsic tasks. Nevertheless, Chiu et al. (2016) and Chiu et al. (2018) report that results from all existing intrinsic evaluation benchmarks in biomedicine fail to reflect how individual models perform in extrinsic tasks. This implies

the better-performing embeddings, as measured by the existing UMNSRS and MayoSRS datasets (intrinsic evaluation), may not perform equivalently well in downstream tasks (extrinsic evaluation).

5.7 Inclusion of External Knowledge

Lexical resources can be used to enrich word embeddings by providing them other sources of linguistic information beyond the distributional statistics obtained from corpora. In recent literature, various methods that leverage knowledge from human-developed and automatically-constructed lexical resources have been proposed. One type of method involves modifying the objectives in the original embedding learning procedures so that they can *jointly learn* both distributional and lexical information. For example, Yu and Dredze (2014) modified the CBOW objective function by introducing semantic constraints (obtained from the paraphrase database (Ganitkevitch et al., 2013)) to train word embeddings that focus on word similarity over word relatedness. Another group of methods incorporate lexical information into the word embeddings as a post-processing procedure. These methods *fine-tune* the pre-trained word embeddings to satisfy linguistic constraints from external resources. The advantage is that they can be applied to any off-the-shelf model without requiring large corpora for (re-) training, as the joint-learning models do. One wide-used fune-tuning approach is *retrofitting* by Faruqui et al. (2015) whose goal is to bring words that are connected via a relation (e.g. synonym) in a given semantic network or lexical resource (i.e. linguistic constraints) closer together in embedding space. For example, Yu et al. (2016) retrofitted the word vector spaces of MeSH terms by using additional linkage information from UMNSRS to improve the embeddings of biomedical concepts. Additionally, building upon this, Lengerich et al. (2018) generalized retrofitting methods by explicitly modelling individual linguistic constraints that are commonly found in health/clinical-related lexicons (e.g. causal-relations between diseases and drugs). In particular, they created two set of embeddings using different settings: one was a Google News pre-trained skip-gram as retrofitted with the lexical frames from FrameNet (Baker et al., 1998) and the other was a wikipage embeddings as retrofitted with the SNOMED-CT ontology. Here, the wikipages were restricted to those that contained SNOMED-CT concepts. They evaluated the embeddings with three Drug-Disease Link Prediction datasets as provided by Godefroy and Potts (2019); Dingwall and Potts (2018) and Tao et al. (2019). While retrofitting, in both settings, improved the embeddings' performance. The better results came when the genres of the retrofitting resources were aligned with the ones in the embedding space (i.e., the SNOMED-CT wikipage embeddings retrofitted with the SNOMED-CT ontology).

5.8 Ensemble of Embeddings

Ensembled word embeddings trained on different corpora allow the prediction model to make use of the different information encoded in each embeddings. Word embeddings pre-trained on general-domain text provide a wide range of vocabulary, while domain-specific word embeddings better represent the properties of in-domain terms. Additionally, the type of medical corpus influences the word embeddings produced. For example, health discussion forum embeddings can better capture the colloquial medical terms used in social media, while PubMed embeddings can better model professional medical terms. Therefore, using an ensemble of general and domain-specific embeddings or an ensemble of embeddings produced from different types of medical corpora are deemed to improve the quality of

embeddings (Belousov et al.; Limsopatham and Collier, 2016; Roberts, 2016).

Apart from combining word embeddings learned from different sources, it is also possible to fuse character and word embeddings. Character embeddings capture the morphological information, such as prefix, suffix and root that make up individual words, whereas word embeddings encode semantic information for individual words, taking into account their contexts. An ensemble of character and word embeddings can thus combine the best of both worlds (Li et al., 2017; Tutubalina and Nikolenko, 2017).

5.9 Incorporating Task-specific Information to Improve Embeddings

For a typical end-to-end NLP system, the quality of the inputted embeddings can be fine-tuned with the addition of task-specific information to boost the performance of prediction. This is done by back-propagating the training errors to the embedding level, as suggested in Collobert et al. (2011). For example, for the task of medical coding in Patel et al. (2017), embeddings were improved with the addition of information from ICD-10 codes.

5.10 Using Contextual Embeddings in Biomedical NLP

An advantage of contextual embeddings, as compared with non-contextual ones like word2vec, is that they can learn dynamic representations for individual words based on the contexts. They are achieving cutting-edge performance in a range of NLP tasks. Biomedical NLP researchers have also demonstrated the importance of transfer learning from pre-trained BERT, where the state-of-the-art performance is obtained by fine-tuning BERT with a large amount of task-/domain-specific data from NER and relation extraction. An example is BioBERT (see Section 4.3).

With their cutting-edge performances, contextual embeddings like BERT and ELMo are now widely used in biomedical NLP. However, since the linguistic properties of biomedical text differ significantly from general English (e.g. it is commonly written in long sentences containing a complex clausal structure full of specific terminologies and acronyms), it is difficult to directly use generic embeddings for biomedical NLP. Hence, there is active research on how to fine-tune contextual embedding methods, particularly from their training setting and model perspectives, to better adapt to biomedical data with optimal performance.

From the training-setting perspective, Beltagy et al. (2019) released SciBERT, which is based on a pre-trained BERT and fine-tuned using scientific publications from the computer science and biomedical domains. SciBERT has outperformed BioBERT in several NER and relation extraction tasks (see Figure 18). Looking for improvement, some changes were applied to SciBERT training to make it better-suit scientific text. First, *ScispaCy*, a science–specific version of spaCy (Neumann et al., 2019), was leveraged to split a document into sentences. Additionally, the authors used the *SentencePiece* library [14] to construct new WordPiece vocabulary for SciBERT rather than using BERT's default vocabulary, as in BioBERT (see Table 16). The results demonstrated room for improvement in applying domain adaption techniques in training contextual embeddings. In the clinical domain, Alsentzer et al. (2019) released two BERT embeddings: one trained on generic clinical text (Clinical BERT) and another on discharge summaries (Discharge Summary BERT). Although the authors demonstrated that using a domain-specific embedding yielded better performance on clinical tasks, they also highlighted that clinical embeddings were not as

[14] `https://github.com/google/sentencepiece`

Task	Dataset	BIOBERT	SCIBERT
NER	BC5CDR	88.85	90.01
	JNLPBA	77.59	77.28
	NCBI-disease	89.36	88.57
REL	ChemProt	76.68	83.64

Figure 18: Performance comparison between BioBERT and SciBERT in NER and relation extraction. Sourced from Beltagy et al. (2019).

	BERT	BioBERT	SciBERT
Corpora	English Wikipedia: 2.5b Book Corpus: 0.8b	English Wikipedia: 2.5b Book Corpus: 0.8b Pubmed: 4.5b PMC:13.5b	Paper from Semantic Scholar Biomedicine: 2.5b Computer Science: 0.6b
Tokenizer	Wordpiece	Wordpiece	ScispaCy: sentence segmentation SentencePiece: tokenization

Table 16: Data and tokenization schemes used by BERT, BioBERT and SciBERT.

effective when trained on de-identification data. In particular, their embeddings were pre-trained on de-identified notes, where all the protected health information (PHI) was removed and replaced with surrogates. For example, a sentence *Mary Smith visited MGH.* would be replaced by *[Patient Name] visited [Hospital].* Consequently, the embeddings might not be well-acquainted for tasks like NER. They suggested future work may consider using synthetic identification data instead (Boag et al., 2018).

From the model perspective, because BERT is trained with the objective of predicting consecutive sentence pair, it can suffer from long-text dependency when the word semantics between non-consecutive sentences is not well-encoded. To address this, XLNet was proposed (Yang et al., 2019). It tries to resolve the long-text dependency issue by modelling on the permutation of sentences rather than the consecutive sentences. When predicting whether patients needed Prolonged Mechanical Ventilation, Huang et al. (2019) reported an improvement from XLNet, compared with BioBERT and Clinical BERT. However, they found that, while XLNet was effective in capturing the temporal nature of long sequences in clinical notes, it was also computationally expensive. Considering that large-scale data are essential for learning effective word features, especially the domain-specific ones, improving model efficiency without negatively impacting embedding quality is thus an active area of research.

5.11 Word Embeddings in Non-Biomedical Domains

The quality of word embeddings can be improved with the addition of domain-specific information. For examples, in the financial domain, Yang et al. (2020) released the FinBERT embeddings. It was created by fine-tuning the generic BERT embeddings with the financial data obtained from posts in Yahoo and Reddit Finance, as well as financial news articles. The authors reported that the fine-tuned embeddings performed better than the generic BERT in financial NLP tasks such as Financial Sentiment Analysis and Question Answering. Similarly, in the legal domain, Chalkidis et al. (2020) reported a better result on legal

document classification when BERT was fine-tuned with legal text sourcing from court cases and contracts.

6 Conclusion

In this paper, we presented a detailed review of word embeddings in biomedical NLP. We began by analyzing and comparing four types of biomedical corpora, including scientific literature, social media text, electronic health records, and knowledge sources. Then, we described four cutting-edge embedding models, categorized by the linguistic features they capture, ranging from morphological to contextual information. Additionally, we also mentioned their real-life applications in several bioinformatic tasks/systems. Following this, we provided an overview of various types of evaluations for word embeddings, including intrinsic and extrinsic approaches. These methods enable researchers to assess, both quantitatively and qualitatively, the different word features captured by individual embedding models, such as word similarity and relatedness.

Later on, we discussed a few novel embedding approaches mentioned in recent literature.These include merging corpora when generating embeddings, including domain knowledge and combining embeddings. Additionally, the problem of missing embeddings for unseen words and misspelled words can be handled using character embeddings like Elmo.

While some issues have, to a certain extent, been addressed, new challenges are emerging. For example, in terms of model interpretability, although embedding models can effectively capture word properties such as syntactic and semantic information, how the embedding quality is affected by the training settings remains unclear. Hence, there is a need for more diversified evaluation resources in order to understand this from different perspectives. The lack of model interpretability makes word embeddings a black box and limits their use in word-type specific tasks like verb classification. Apart from this, while general-domain embeddings provide a large coverage of vocabularies, they are not necessarily applicable to domain-specific tasks (e.g. in biomedicine) that are linguistically distinct from general English. To achieve maximum benefit when using word embeddings for biomedical NLP tasks, they need to be produced and evaluated using in-domain text.This opens the door to a potential research avenue in domain-adaptation methodologies for extending the generalizability of embeddings.

Viewed as a whole, the overview presented by this paper demonstrates the practical application of word embeddings in biomedicine. In particular, our work shows that neural word embeddings can be used to benefit biomedical NLP in many ways.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, 2019.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*, 2016.

Apache Software Foundation. openNLP Natural Language Processing Library, 2014. URL `http://opennlp.apache.org/`. http://opennlp.apache.org/.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics, 1998.

Simon Baker, Anna Korhonen, and Sampo Pyysalo. Cancer hallmark text classification using convolutional neural networks. *BioTxtM 2016*, page 1, 2016.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226, 2009.

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurélie Névéol, Mariana Neves, Felipe Soares, et al. Findings of the wmt 2019 biomedical translation shared task: Evaluation for medline abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, 2019.

Andrew L Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin Weber, Nathan Palmer, Xu Shi, Tianxi Cai, and Isaac S Kohane. Clinical concept embeddings learned from massive sources of multimodal medical data. In *Pacific Symposium on Biocomputing*, volume 25, pages 295–306, 2020.

M Belousov, W Dixon, and G Nenadic. Using an ensemble of linear and deep learning models in the smm4h 2017 medical concept normalization task in: Proceedings of the second workshop on social media mining for health research and applications workshop co-located with the american medical informatics association annual symposium (amia 2017); 2017: 54–58.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3606–3611, 2019.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, pages 1137–1155, 2003.

Willie Boag and Hassan Kané. Awe-cm vectors: Augmenting word embeddings with a clinical metathesaurus. *arXiv preprint arXiv:1712.01460*, 2017.

Willie Boag, Tristan Naumann, and Peter Szolovits. Towards the creation of a large corpus of synthetically-identified clinical notes. *arXiv preprint arXiv:1803.02728*, 2018.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

on behalf of the BNC Consortium Bodleian Libraries, University of Oxford. The british national corpus, version 3 (bnc xml edition). `http://www.natcorp.ox.ac.uk/`, May 2007.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. ISSN 2307-387X.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pages 12–58, 2014.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, 2016.

Fabienne Braune and Alexander Fraser. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 81–89. Association for Computational Linguistics, 2010.

Elliot G Brown, Louise Wood, and Sue Wood. The medical dictionary for regulatory activities (meddra). *Drug safety*, 20(2):109–117, 1999.

Xiangrui Cai, Jinyang Gao, Kee Yuan Ngiam, Beng Chin Ooi, Ying Zhang, and Xiaojie Yuan. Medical concept embedding with time-aware attention. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 3984–3990, 2018.

Shaosheng Cao, Wei Lu, Jun Zhou, and Xiaolong Li. cw2vec: Learning chinese word embeddings with stroke n-gram information. In *AAAI*, pages 5053–5061, 2018.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*, 2020.

Kaytlin Chaloner and Alfredo Maldonado. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32, 2019.

Zhengping Che, Yu Cheng, Shuangfei Zhai, Zhaonan Sun, and Yan Liu. Boosting deep learning risk prediction with generative adversarial networks for electronic health records. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 787–792. IEEE, 2017.

Hong-You Chen, Sz-Han Yu, and Shou-De Lin. Glyph2vec: Learning chinese out-of-vocabulary word embedding from glyphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2865–2871, 2020.

Zhiwei Chen, Zhe He, Xiuwen Liu, and Jiang Bian. Evaluating semantic relations in neural word embeddings with biomedical and general domain knowledge bases. *BMC medical informatics and decision making*, 18(2):65, 2018.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. How to train good word embeddings for biomedical nlp. In *Proceedings of the 15th workshop on biomedical natural language processing*, pages 166–174, 2016.

Billy Chiu, Sampo Pyysalo, Ivan Vulić, and Anna Korhonen. Bio-simverb and bio-simlex: wide-coverage evaluation sets of word similarity in biomedicine. *BMC bioinformatics*, 19 (1):33, 2018.

Billy Chiu, Olga Majewska, Sampo Pyysalo, Laura Wey, Ulla Stenius, Anna Korhonen, and Martha Palmer. A neural classification method for supporting the creation of bioverbnet. *Journal of biomedical semantics*, 10(1):2, 2019.

Youngduck Choi, Chill Yi-I Chiu, and David Sontag. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings*, 2016: 41, 2016.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167, 2008.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. A neural network multi-task learning approach to biomedical named entity recognition. *BMC bioinformatics*, 18 (1):368, 2017.

Chris Culnane, Benjamin IP Rubinstein, and Vanessa Teague. Health data in an open world. *arXiv preprint arXiv:1712.05627*, 2017.

Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Peter Bruza. Medical semantic similarity with a neural language model. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 1819–1822, 2014.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Nicholas Dingwall and Christopher Potts. Mittens: an extension of glove for learning domain-specialized representations. In *Proceedings of the 2018 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 212–217, 2018.

Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. Accountability of ai under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*, 2017.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, pages 3519–3530, 2016.

Sebastien Dubois and Nathanael Romano. Learning effective embeddings from medical notes. *arXiv preprint arXiv:1705.07025*, 2017.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. Retrofitting word vectors to semantic lexicons. In *Proc. of NAACL*, 2015.

Sergey Feldman, Waleed Ammar, Kyle Lo, Elly Trepman, Madeleine van Zuylen, and Oren Etzioni. Quantifying sex bias in clinical studies at scale with automated data extraction. *JAMA network open*, 2(7):e196700–e196700, 2019.

Roger B Fillingim, Christopher D King, Margarete C Ribeiro-Dasilva, Bridgett Rahim-Williams, and Joseph L Riley III. Sex, gender, and pain: a review of recent clinical and experimental findings. *The journal of pain*, 10(5):447–485, 2009.

Samuel G Finlayson, Paea LePendu, and Nigam H Shah. Building the graph of medicine from millions of clinical narratives. *Scientific data*, 1:140032, 2014.

Peter W Foltz. Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, & Computers*, 28(2):197–202, 1996.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, 2013.

Martin Gerner, Goran Nenadic, and Casey M Bergman. Linnaeus: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):85, 2010.

Bruno Godefroy and Christopher Potts. Modeling drug–disease relations with linguistic and knowledge graph constraints. Ms., Roam Analytics and Stanford University. arXiv:1904.00313, 2019.

Gene H Golub and Christian Reinsch. Singular value decomposition and least squares solutions. In *Linear Algebra*, pages 134–151. Springer, 1971.

Rodrigo Rafael Villarreal Goulart, Vera Lúcia Strube de Lima, and Clarissa Castellã Xavier. A systematic review of named entity recognition in biomedical texts. *Journal of the Brazilian Computer Society*, 17(2):103–116, 2011.

Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

Johannes Hellrich, Simon Clematide, Udo Hahn, and Dietrich Rebholz-Schuhmann. Collaboratively annotating multilingual parallel corpora in the biomedical domain—some mantras. In *LREC*, pages 4033–4040. Citeseer, 2014.

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 123–129, 2018.

William H Herman and Robert M Cohen. Racial and ethnic differences in the relationship between hba1c and blood glucose: implications for the diagnosis of diabetes. *The Journal of Clinical Endocrinology & Metabolism*, 97(4):1067–1072, 2012.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics, 2012.

Jian Huang, Keyang Xu, and VG Vinod Vydiswaran. Analyzing multiple medical corpora using word embedding. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 527–533. IEEE, 2016.

Kexin Huang, Abhishek Singh, Sitong Chen, Edward T Moseley, Chih-ying Deng, Naomi George, and Charlotta Lindvall. Clinical xlnet: Modeling sequential clinical notes and predicting prolonged mechanical ventilation. *arXiv preprint arXiv:1912.11975*, 2019.

Julia Ive, Aurélien Max, François Yvon, and Philippe Ravaud. Diagnosing high-quality statistical machine translation using traces of post-edition operations. In *Proceedings of the LREC 2016 workshop on translation evaluation—from fragmented tools and data sets to an integrated ecosystem, Portorož, Slovenia*, pages 55–62, 2016.

Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395, 2012.

Qiao Jin, Bhuwan Dhingra, William W Cohen, and Xinghua Lu. Probing biomedical embeddings from language models. *NAACL HLT 2019*, page 82, 2019.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

Pentti Kanerva, Jan Kristoferson, and Anders Holst. Random indexing of text samples for latent semantic analysis. In *In Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 103–6. Erlbaum, 2000.

Amitabha Karmakar. Classifying medical notes into standard disease codes using machine learning. *arXiv preprint arXiv:1802.00382*, 2018.

Aris Kosmopoulos, Ion Androutsopoulos, and Georgios Paliouras. Biomedical semantic indexing using dense word vectors in bioasq. *Journal Of Biomedical Semantics*, 2015.

Michael Krauthammer and Goran Nenadic. Term identification in the biomedical literature. *Journal of biomedical informatics*, 37(6):512–526, 2004.

Andrei Kutuzov and Pierre Lison. Redefining context windows for word embedding models: An experimental study. In *Proceedings of the 58th Conference on Simulation and Modelling*, pages 284–288. Linköping University Electronic Press, 2017.

Fethi Lamraoui and Philippe Langlais. Yet another fast, robust and open source sentence aligner. time to reconsider sentence alignment. *XIV Machine Translation Summit*, 2013.

Thomas K Landauer and Susan T Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

Hoang-Quynh Le, Duy-Cat Can, Sinh T Vu, Thanh Hai Dang, Mohammad Taher Pilehvar, and Nigel Collier. Large-scale exploration of neural relation classification architectures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2266–2277, 2018.

Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 09 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682. URL `https://doi.org/10.1093/bioinformatics/btz682`.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Ben Lengerich, Andrew Maas, and Christopher Potts. Retrofitting distributional embeddings to knowledge graphs with functional relations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2423–2436. Association for Computational Linguistics, 2018. URL `http://aclweb.org/anthology/C18-1205`.

Paea LePendu, Srinivasan V Iyer, Anna Bauer-Mehren, Rave Harpaz, Jonathan M Mortensen, Tanya Podchiyska, Todd A Ferris, and Nigam H Shah. Pharmacovigilance using clinical notes. *Clinical pharmacology & therapeutics*, 93(6):547–555, 2013.

Ulf Leser and Jörg Hakenberg. What makes a gene name? named entity recognition in the biomedical literature. *Briefings in bioinformatics*, 6(4):357–369, 2005.

Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, 2014a.

Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 171–180, 2014b.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185, 2014c.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. Exploiting bert for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 34–41, 2019.

Zhenzhen Li, Qun Zhang, Yang Liu, Dawei Feng, and Zhen Huang. Recurrent neural networks with specialized word embedding for chinese clinical named entity recognition. In *CEUR Workshop Proceedings*, volume 1976, pages 55–60, 2017.

Nut Limsopatham and Nigel Collier. Modelling the combination of generic and target domain embeddings in a convolutional neural network for sentence classification. Association for Computational Linguistics, 2016.

Wang Ling, Chris Dyer, Alan W. Black, and Isabel Trancoso. Two/too simple adaptations of word2vec for syntax problems. In *HLT-NAACL*, 2015a.

Wang Ling, Yulia Tsvetkov, Silvio Amir, Ramon Fermandez, Chris Dyer, Alan W Black, Isabel Trancoso, and Chu-Cheng Lin. Not all contexts are created equal: Better word representations with variable attention. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1367–1372. Association for Computational Linguistics, 2015b. doi: 10.18653/v1/D15-1161. URL `http://aclweb.org/anthology/D15-1161`.

Carolyn E Lipscomb. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3721–3731, 2019.

Henry J Lowe, Todd A Ferris, Penni M Hernandez, and Susan C Weber. Stride–an integrated standards-based translational research informatics platform. In *AMIA Annual Symposium Proceedings*, volume 2009, page 391. American Medical Informatics Association, 2009.

Minh-Thang Luong, Richard Socher, and Christopher D Manning. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113, 2013.

Christopher D Manning, Hinrich Schütze, and Prabhakar Raghavan. *Introduction to information retrieval*. Cambridge university press, 2008.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. URL `http://www.aclweb.org/anthology/P/P14/P14-5010`.

Bridget T McInnes and Ted Pedersen. Evaluating semantic similarity and relatedness over the semantic grouping of clinical term pairs. *Journal of biomedical informatics*, 54:329–336, 2015.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119, 2013b.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, 2013c.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38 (11):39–41, 1995.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. Scispacy: Fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*, 2019.

Mariana Neves. A parallel collection of clinical trials in portuguese and english. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 36–40, 2017.

Mariana Neves, Antonio Jimeno Yepes, and Aurélie Névéol. The scielo corpus: a parallel corpus of scientific publications for biomedicine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2942–2948, 2016.

Serguei VS Pakhomov, Ted Pedersen, Bridget McInnes, Genevieve B Melton, Alexander Ruggieri, and Christopher G Chute. Towards a framework for developing semantic relatedness reference standards. *Journal of biomedical informatics*, 44(2):251–265, 2011.

Kevin Patel, Divya Patel, Mansi Golakiya, Pushpak Bhattacharyya, and Nilesh Birari. Adapting pre-trained word embeddings for use in medical coding. In *BioNLP 2017*, pages 302–306, 2017.

Ted Pedersen, Serguei VS Pakhomov, Siddharth Patwardhan, and Christopher G Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics*, 40(3):288–299, 2007.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, volume 14, pages 1532–1543, 2014.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

Bruno Pouliquen and Christophe Mazenc. Coppa, clir and tapta: three tools to assist in overcoming the pat-ent language barrier at wipo. *Proceedings of the 13th Machine Translation Summit*, pages 24–30, 2011.

Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski, and Sophia Ananiadou. Distributional semantics resources for biomedical text processing. *Proceedings of LBM*, 2013.

Marek Rei, Gamal Crichton, and Sampo Pyysalo. Attending to characters in neural sequence labeling models. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 309–318, 2016a.

Marek Rei, Gamal K. O. Crichton, and Sampo Pyysalo. Attending to characters in neural sequence labeling models. In *COLING*, 2016b.

Kirk Roberts. Assessing the corpus size vs. similarity trade-off for word embeddings in clinical nlp. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 54–63, 2016.

Helmut Schmid. Probabilistic part-ofispeech tagging using decision trees. In *New methods in language processing*, page 154, 2013.

Hinrich Schütze and Jan Pedersen. A vector model for syntagmatic and paradigmatic relatedness. In *Proceedings of the 9th Annual Conference of the UW Centre for the New OED and Text Research*, pages 104–113. Citeseer, 1993.

Michael Q Stearns, Colin Price, Kent A Spackman, and Amy Y Wang. Snomed clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662. American Medical Informatics Association, 2001.

Pontus Stenetorp, Hubert Soyer, Sampo Pyysalo, Sophia Ananiadou, and Takashi Chikayama. Size (and domain) matters: Evaluating semantic word space representations for biomedical text. *Proceedings of SMBM'12*, 2012.

Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2019.

Yifeng Tao, Bruno Godefroy, Guillaume Genthial, and Christopher Potts. Effective feature representation for clinical text concept extraction. *NAACL Workshop on Clinical Natural Language Processing*, 2019.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*, 2019.

Fei Tian, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu. A probabilistic model for learning multi-prototype word embeddings. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 151–160, 2014.

Jörg Tiedemann. News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248, 2009.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218, 2012.

Yota Toyama, Makoto Miwa, and Yutaka Sasaki. Utilizing visual forms of japanese characters for neural review classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 378–382, 2017.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. Evaluation of word vector representations by subspace alignment. In *Proc. of EMNLP*, 2015.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics, 2010.

Peter D Turney. Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research*, pages 533–585, 2012.

Elena Tutubalina and Sergey Nikolenko. Combination of deep recurrent neural networks and conditional random fields for extracting adverse drug reactions from user reviews. *Journal of healthcare engineering*, 2017, 2017.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4*, 292:247, 2007.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Ellen M Voorhees and William R Hersh. Overview of the trec 2012 medical records track. In *TREC*, 2012.

PJTM Vossen. Eurowordnet: a multilingual database for information retrieval. In *Proceedings of the DELOS workshop on Cross-language Information Retrieval, March 5-7, 1997 Zurich*. Vrije Universiteit, 1997.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. Hyperlex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics*, 43(4):781–835, 2017.

Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. A comparison of word embeddings for the biomedical natural language processing. *Journal of biomedical informatics*, 87:12–20, 2018.

Katharina Wäschle and Stefan Riezler. Analyzing parallelism and domain similarities in the marec patent corpus. In *Information Retrieval Facility Conference*, pages 12–27. Springer, 2012.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, 2018.

Yi Yang, Mark Christopher Siy UY, and Allen Huang. Finbert: A pretrained language model for financial communications. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20) Special Track on AI in FinTech*, 2020.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764, 2019.

Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Karin Verspoor, Ondřej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, et al. Findings of the wmt 2017 biomedical translation shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 234–247, 2017.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.

Mo Yu and Mark Dredze. Improving lexical embeddings with semantic knowledge. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 545–550, 2014.

Zhiguo Yu, Trevor Cohen, Byron Wallace, Elmer Bernstam, and Todd Johnson. Retrofitting word vectors of mesh terms to improve semantic similarity measures. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 43–51, 2016.

Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. Hurtful words: quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120, 2020.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):1–9, 2019.

Mengnan Zhao, Aaron J Masino, and Christopher C Yang. A framework for developing and evaluating word embeddings of drug-named entity. In *Proceedings of the BioNLP 2018 workshop*, pages 156–160, 2018.

Henghui Zhu, Ioannis C Paschalidis, and Amir M Tahmasebi. Clinical concept extraction with contextual word embedding. In *NIPS Machine Learning for Health Workshop*, 2018.

Yongjun Zhu, Erjia Yan, and Fei Wang. Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. *BMC medical informatics and decision making*, 17(1):95, 2017.