Journal of Biomedical Informatics 104 (2020) 103396

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Evaluating sentence representations for biomedical text: Methods and experimental results

Noha S. Tawfik[a,b,*], Marco R. Spruit[b]

[a] Computer Engineering Department, College of Engineering, Arab Academy for Science, Technology, and Maritime Transport (AAST), 1029 Alexandria, Egypt
[b] Department of Information and Computing Sciences, Utrecht University, 3584 CC Utrecht, the Netherlands

ARTICLE INFO

Keywords:
BioNLP
Text representation
Sentence embeddings
Language model

ABSTRACT

Text representations ar one of the main inputs to various Natural Language Processing (NLP) methods. Given the fast developmental pace of new sentence embedding methods, we argue that there is a need for a unified methodology to assess these different techniques in the biomedical domain. This work introduces a comprehensive evaluation of novel methods across ten medical classification tasks. The tasks cover a variety of BioNLP problems such as semantic similarity, question answering, citation sentiment analysis and others with binary and multi-class datasets. Our goal is to assess the transferability of different sentence representation schemes to the medical and clinical domain. Our analysis shows that embeddings based on Language Models which account for the context-dependent nature of words, usually outperform others in terms of performance. Nonetheless, there is no single embedding model that perfectly represents biomedical and clinical texts with consistent performance across all tasks. This illustrates the need for a more suitable bio-encoder. Our MedSentEval source code, pretrained embeddings and examples have been made available on GitHub.

1. Introduction

In the past few years, neural network-based distributional representations of text such as word embeddings have been shown highly effective in solving multiple NLP problems. There are many different types of word embeddings [1]. However, they all have the same purpose: to generate low-dimensional vector representations of words. They can encode important syntactic properties of words efficiently, and are able to capture semantic similarity among words as mathematical similarities between their vectors. Similarly, sentence embeddings are numerical representations of sentences, which are often derived from word embeddings. Nonetheless, the last two years witnessed a rise of different supervised and unsupervised approaches towards learning representations of sequences of words, such as sentences or paragraphs. They can identify the order of words within a sentence and hence capture more context. The developed sentence representations extend the success of earlier word vector-based approaches with interesting results and increasing potential across different tasks.

The progress in machine learning has given scientists the unprecedented opportunity to extract valuable information from biomedical data. With the increasing availability of unstructured textual data in the biomedical domain in forms of clinical trials, research articles, electronic health records, and patient-authored texts, the use of text mining techniques is becoming increasingly more important. The importance of word embeddings in Biomedical Natural Language Processing (BioNLP) becomes evident by looking at the number of recent researches in the field. These embeddings have been commonly leveraged as feature input for several BioNLP tasks. Word-level embeddings have been studied extensively in the biomedical domain [2–4]. On the other hand, the analysis of sentence-level representations has been much more limited to a few scattered works [5,6] and there is a lack of a full analysis of embedding techniques on common grounds.

Motivated by [7,8], in their efforts to evaluate sentence representations in a fair and structured approach, this paper aims at replicating their evaluations in domain-specific settings. More specifically, we assess the ability of existing sentence representation techniques to capture the rich and complex semantics of clinical sentences. We focus on what are arguably the state-of-art techniques in embedding sentences known for achieving high performance in general NLP tasks. Throughout our analysis, we test and compare several sentence embedding methods trained on general, medical and clinical data. Our evaluations include multiple classification problems related to the clinical and biomedical domain spanning different linguistic tasks. We discuss the strengths and weaknesses of the different techniques in encoding domain-specific aspects of clinical sentences. To our knowledge, no similar evaluation exists for the biomedical domain. This paper

* Corresponding author at: Computer Engineering Department, College of Engineering, Arab Academy for Science, Technology, 1029 Alexandria, Egypt.
E-mail addresses: noha.abdelsalam@aast.edu, n.s.tawfik@uu.nl (N.S. Tawfik), m.r.spruit@uu.nl (M.R. Spruit).

https://doi.org/10.1016/j.jbi.2020.103396
Received 12 July 2019; Received in revised form 22 January 2020; Accepted 24 February 2020
Available online 06 March 2020
1532-0464/

is organized as follows: In Section 2, we give a background of word and sentence embeddings. In Section 3, we provide details of all ten tasks included in our evaluations. We also describe the experimental settings of the sentence embedding models implemented. In Sections 4 to 6, we illustrate the results obtained and draw corresponding conclusions accordingly.

2. Background

Words representations are inspired by the concept of distributional semantic models that hypothesize that word meanings could be inferred by the company they keep [9]. While the concept is old, its recent popularity could be traced back to the work of Bengio et al. on natural language modeling through a neural probabilistic model [10]. Among the first approaches to embed words based on neural networks is the Word2vec algorithm proposed by Mikolov et al. [11]. The model is a shallow, three-layered neural network that uses unsupervised learning to determine the semantic and syntactic meanings of a word based on adjacent words denoted as context. It offers two variations: Continuous Bag of Words (CBOW) and Skip-gram. The first learns the representations by predicting the target word based on its context words while skip-gram inverts contexts and targets, and tries to predict each context word from its target word, rather than predicting the target word itself. Global Vector word representations (GloVe) [12] is another prominent method that enabled efficient unsupervised training of dense word representations and straightforward integration into NLP tasks. GloVe is a count-based model, as opposed to Word2Vec that is considered a predictive model. Count-based models utilize the word distribution statistics of the corpus effectively. It constructs a global word-word co-occurrence matrix and applies matrix factorization to learn lower dimension embeddings, where each row is some vector representation for each word. FastText is a recent addition to prediction-based models, proposed by Facebook, for learning word embeddings over large datasets. Its architecture is similar to the skip-gram model, but it includes a significant improvement that accounts for the morphological properties of words through the learning process [13]. Embeddings are produced by combining the n-gram embeddings for all the n-grams characters in a word. The main advantage over prior techniques is its ability to predict out-of-vocabulary (OOV) words as a result of its subword representations. The above models are static, context independent and do not account for polysemy. In other words, the model outputs only one vector for each word regardless of the word position in the sentence, or the context in which it appears [14].

On the other hand, embeddings based on Language Models (LM) dynamically change so they can discriminate among different meanings of a word. Language modeling serves as an unsupervised pre-training stage, where learning is independent of the main task, on a large unlabeled or differently-labeled text corpus. It can generate the next word in a sentence with knowledge of previous words. These resulting embeddings are the internal states of deep neural networks in a monolingual or a bilingual language modeling setting. Among the first attempts to generate context-sensitive representations is Context2vec [15]. The model represents the context of a target word by extracting the output embedding of a multi-layer perceptron built on top of a bidirectional Long short-term memory (LSTM) language model. Other examples include Embeddings from Language Models (ELMo) [16], Bidirectional Encoder Representations from Transformers (BERT) [17], Universal Language Model Fine-tuning (ULMFiT) [18] and the Pooled Contextualized Embeddings from Flair toolkit [19].

ELMo Vectors are also computed on top of two-layer bidirectional Language Models (biLMs) with character convolution. Using CNNs, each vector is built upon the characters that compose the underlying words. BERT is different from ELMo primarily because it targets a different training objective; it uses masked language modeling instead of traditional LM. It overcomes ELMo limitations by including left and right contexts simultaneously when representing words. BERT replaces words in a sentence randomly and inserts a “masked” token. The transformer generates predictions for the masked words based on left and right unmasked neighbors. FLAIR contextualized embeddings are word-level embeddings that were shown effective in the sequence labeling task. Input sentences are modeled as distributions over sequences of characters to a bidirectional character-level language model. The neural model is pre-trained on large unlabeled corpora, and internal character states are used to compute the output word embeddings.

To obtain sentence embeddings on top of word embeddings, a simple Bag-of-Words (BoW) inspired method could be applied by computing the mean of the vector embeddings for the words in a sentence. Alternatively, more advanced and sophisticated methods such as Sent2Vec and Smooth Inverse Frequency (SIF) could be employed. In the Sent2Vec paradigm, a sentence embedding is defined as the average of the source word embeddings of its constituent words [20]. The method is furthermore augmented by learning source embeddings for unigrams and n-grams of words present in each sentence, and averaging the n-gram embeddings along with the words. In contrast, SIF adds a weighting function to word embeddings, which down-weights common words [21].

However, despite the improved performance achieved using sophisticated methods, the main limitation of conventional embeddings at the word-level is the negligence of the overall sentence structure. Nevertheless, some scholars argue that regardless of the potential information loss, word embeddings are still able to represent sentence meanings efficiently [21,23].

Alternatively, there have been efforts to generate dedicated sentence embeddings through unsupervised training. The Skip-Thought model extends the original skip-gram algorithm from words to sentences [24]. It predicts neighbour sentences or phrases for a given sentence using a recurrent neural network. It follows an encoder-decoder model where first a sentence is encoded into a vector through a Gated Recurrent Units (GRU) or LSTM architecture, and that representation is decoded into surrounding text. It adopts a vocabulary expansion scheme that makes use of pre-trained embeddings to learn embeddings of new non-encountered words.

In contrast, Facebook introduces inferSent [25], a supervised learning methodology for sentence encoding. Their work provides solutions to two critical concerns: the best training task and network architecture to obtain a universal sentence representation model. Their findings indicate that detecting natural language inference is the most suitable for transfer learning to other NLP tasks. This is attributed to the semantic nature of the task and the availability of a very large corpus such as the Stanford Natural Language Inference (SNLI) that consists of 570k humanly generated English sentence pairs, manually labeled. Moreover, they experiment with seven different architectures including standard recurrent encoders with either LSTM or GRU, concatenation of last hidden states of forward and backward GRU, Bi-directional LSTMs (BiLSTM) with either mean or max pooling, self-attentive networks, and hierarchical convolutional networks. They conclude that the BiLSTM with the max-pooling operation performs best on both SNLI and transfer tasks. Google also published a sentence encoder known as Universal Sentence Embeddings (USE) [26]. It is referred to as “universal” since, in theory, it is supposed to encode general properties of sentences given the large size of datasets it is trained on. The multi-task learning encoder uses several annotated and unannotated datasets for training. It has two variants of the encoding architectures. The Transformer model is designed for higher accuracy, but the encoding requires more memory and computational time. The Deep Averaging Network (DAN) model, on the other hand, is designed for speed and efficiency, and some accuracy is compromised. When integrated with any downstream task, USE should be able to represent sentences efficiently without the need for any domain-specific knowledge. This is a great advantage when limited training resources are available for specific tasks.

We include GloVe, FastText, ELMo, BERT, Flair, Infersent, and USE embeddings in our evaluations as we believe that they successfully represent all different techniques previously discussed.

3. Methods

In this paper, we propose *MedSentEval*,[1] an embedding evaluation toolkit designed for the medical domain. It can compute, evaluate, and classify pre-trained sentence embeddings for several BioNLP tasks. The proposed toolkit heavily makes use of SentEval.[2] a general evaluation protocols toolkit. This section describes the included tasks and gives further details on the pre-trained models supported in our toolkit and the evaluation procedures for each task.

3.1. Evaluation tasks

One of the main challenges in the biomedical NLP domain is the availability of benchmark corpora for evaluation. The creation of a dataset faces many barriers such as privacy issues for patient data protection due to the sensitive nature of data, the inefficacy of using crowd-sourcing platforms to annotate data and the need to rely on domain experts which endures more costs. Despite the limitations mentioned above, there have been efforts in creating datasets. However, they are relatively small in size and mostly focus on information extraction. In this paper, we gathered BioNLP datasets that are suitable for classification problems and cover a variety of NLP tasks, including binary and multi-class classification. Below, we provide a brief description of each dataset grouped by types of tasks. Table 1 summarizes the details of each of the datasets used in our experiments and provide examples.

Textual Entailment (TE). TE is an important task in the NLP domain. Given two snippets of text, Text (T) and Hypothesis (H), the TE recognition determines if the meaning of H can be inferred from that of T [27].

The medical natural language inference benchmark dataset *MedNLI* is a source of biomedical TE data derived from clinical notes [28,29]. Its creation process is similar to the creation of the gold-standard SNLI dataset with adaptation to the clinical domain.

Expert annotators were presented with 4638 premises extracted from the MIMIC-III database [30] and were asked to write three hypotheses with true, false, and neutral descriptions of each premise. The final dataset comprises 14,049 sentence pairs divided into 11,232, 1,395 and 1,422 for training, development and testing, respectively.

Recognizing Question Entailment *RQE*, tackles the problem of finding duplicate questions by labeling questions based on their similarity [31,32]. Extending the former textual entailment definition, the authors define question entailment as “Question A entails Question B if every answer to B is also a correct answer to A exactly or partially.” The *RQE* dataset is specifically designed to find the most similar frequently asked question (FAQ) to a given question. The training set was constructed from the questions provided by family doctors on the National Library of Medicine (NLM) platform resulting in 8,588 question pairs where 54.2% are positive pairs. For the test set, two sources of questions were used: validated questions from the NLM collections and FAQs retrieved from the National Institutes of Health (NIH) website. The test set corpus includes 302 pairs of questions, with 42.7% pairs positively labeled.

Sentence classification. In recent years, there has been a substantial increase in the number of scientific publications in the biomedical domain with valuable evidence-based medicine guidelines. Consequently, many NLP methods were deployed to automate or semi-automate the analysis of large medical literature databases such as PubMed. Both datasets included in this category use randomized controlled trials (RCT) as a source of data. The *PICO* dataset is curated from abstracts of RCTs available in the PubMed repository [33]. It maps each abstract sentence into the known Participants, Intervention, Comparison, and Outcome (PICO) elements [34,35]. Moreover, the authors extended the original PICO framework and added three additional categories: aim, method, and results. The annotated data include 24,668 abstracts; each sentence was assigned to a category according to a predefined keyword list compiled manually. The final dataset contains approximately 257 K, 31 K, and 30 K sentences for training, testing, and validation. The *PUBMED20K* corpus is designed for sequential sentence classification of RCTs textual data. Abstracts’ sentences are labeled according to their role in the abstract into background, objective, method, result, or conclusion [36,37]. The data collection process was limited to randomized controlled trial abstracts with a structured format. The dataset is large in size with around 180 K sentences for training, 30 K sentences for validation, and another 30 K sentences for testing with a total of 20,000 abstracts.

Sentiment analysis. Reproducibility is very common in biomedical research where many studies try to replicate earlier work. Scientists express their opinions in many different ways, specifically when citing other studies. The citation sentiment analysis corpus *CitationSA* [38,39] is the first of its kind in the biomedical domain. It includes the discussion section of 285 randomly selected clinical trial abstracts. The annotation scheme did not consider the correctness of the published claims and polarity was assigned on the citation level according to context and not at the sentence level. Two medical annotators labeled sentences according to the former scheme, and a third annotator was involved in case of disagreement. The total number of citation sentences included in the dataset is 4,182 citations. It is also important to include a dataset that represents patient opinions on health-related topics. Patient authored texts usually mix medical terminologies with informal language. *VaccineSA* is a collection of English tweets that includes HPV vaccination keywords [40,41]. The tweets were classified according to their content and polarity into positive, negative, neutral, and unrelated. The negative category is further divided based on the negative concern such as safety, efficacy, cost, resistant, or other. The dataset originally contained 3984 tweets, however, when we re-collected the tweets, only 1853 were available.

Question answering. This task is a longstanding problem extensively studied in the past years and is currently gaining interest in the biomedical domain. The BioASQ challenge [42,43] targets different stages of the question answering process, ranging from the retrieval of relevant concepts and articles to the generation of natural language answers. For the classification task, our interest is in the second phase of task B where BioASQ released questions from benchmark datasets created by a group of biomedical experts. The questions are accompanied by text snippets extracted from relevant PubMed and PMC articles. Four question types are included in the challenge: yes/no, factoid, list, and summary questions, we experiment with the yes/no category only. The BioASQ 6b-task dataset includes 612 question-answer pairs for training and 130 pairs for testing.

We also add the Bio-Contradiction *BioC* dataset to the evaluation. Although it was originally built to detect contradictions among published findings, the dataset structure is also suitable for the question answering task [44,45]. It is organized into 24 PICO questions related to cardiovascular disease generated from systematic reviews. Two annotators were asked separately to curate all research abstracts of studies referenced in the systematic review and extract one sentence per abstract that answers the question. Sentences are labeled *YES* if they positively answer the question and NO otherwise. The corpus has a total of 259 question-answer pairs, out of which 180 are labeled *YES* and 79 labeled *NO*.

Semantic Text Similarity (STS). Different than the former classification tasks, the goal of this task is to measure the relatedness of two sentences and compare it with a human-labeled similarity score.

[1] https://github.com/nstawfik/MedSentEval.
[2] https://github.com/facebookresearch/SentEval.

Table 1
Evaluation datasets description and examples.

Dataset	Task	Source	Example	Label
MedNLI	Textual Entailment	Patient records	*H1:During hospitalization, patient became progressively more dyspnic requiring BiPAP and then a NRB* *P2:The patient is on room air*	Contradiction
RQE	Question Entailment	Doctor questions	*Q1: What should I do with this patient whose biopsy report shows carcinoma in situ of the vulva?* *Q2: What to do with this patient, biopsy shows carcinoma in situ of the vulva?*	True
PUBMED20K	Sentence Classification	Medical articles	*Text:Transient intraocular pressure elevation and cataract progression occurred.*	Background
PICO	Sentence Classification	Medical articles	*Text: Classes included CRC survivors and people with CVD.*	Intervention
PatientSA	Sentiment Analysis	Patient tweets	*Text: Don't forget to also vaccinate your sons. It is potentially even more important. #HPV #vaccineswork*	Positive
CitationSA	Sentiment Analysis	Medical articles	*Text: Patrek et al. [C] examined 13 factors influencing fluid drainage.*	Neutral
BioASQ	Question Answering	Medical articles	*Q:Is osteocrin expressed exclusively in the bone?* *A:Evolution of Osteocrin as an activity-regulated factor in the primate brain.*	No
BioC	Question Answering	Medical	*Q:In women with pre-eclampsia, is mutation in renin-angiotensin gene associated with pre-eclampsia?* *A:The variants(A– > C) of 1166 polymorphism site of AT1RG predisposes increased risk of PIH.*	Yes
C-STS	Semantic Similarity	Patient records	S1: Use information was down loaded from the patient's PAP device and reviewed with the patient. S2:I discussed the indications, contraindications and side effects of doxycycline with the patient.	0.5
BIOSSES	Semantic Similarity	Medical articles	*S1: The oncogenic activity of mutant Kras appears dependent on functional Craf.* *S2: Oncogenic KRAS mutations are common in cancer.*	1

Clinical Semantic Textual Similarity *ClinicalSTS* was published as part of the shared task at the 2018 BioCreative/OHNLP challenge [46,47]. This challenge was organized to investigate the STS problem in the clinical domain following the lead of the original SemEval STS shared tasks. The dataset is a randomly annotated subset of the MedSTS dataset that consists of a total of 174,629 sentences. The dataset was collected from patients records at the Mayo Clinic's clinical data warehouse. Three surface lexical similarities were employed to find candidate pairs: Ratcliff/Obershelp pattern matching algorithm, cosine similarity, and Levenshtein distance. More details on the original MedSTS dataset construction could be found in [48]. For ClinicalSTS, the sentence pairs were annotated independently by two clinical experts who scored each pair based on their semantic equivalence. Scores ranged from 0 to 5, where 0 denotes complete dissimilarity between sentences. The final similarity value was set to the average of both annotators' scores. The dataset includes 1,068 sentence pairs with 70% (750 sentence pairs) and 30% (318 sentence pairs) for training and testing, respectively. All included sentences are de-identified sentences as all protected health information (PHI) was removed through a frequency filtering approach followed by a manual check. The second corpus, the biomedical sentence similarity estimation *BIOSSES* corpus [49,50] comprises 100 sentence pairs. All sentences are extracted from the biomedical summarization track of the Text Analysis Conference (TAC). The subset includes sentences from biomedical articles with a citation mention to a reference article. Similar to *ClinicalSTS*, the dataset creators followed the SemEval guidelines for annotations with five experts giving 0 to 4 score values to sentence pairs to indicate no relation (0) or equivalent (4).

3.2. Embedding methods

GloVe We use the pre-trained embeddings consisting of 2.2 million vocabulary words available at https://nlp.stanford.edu/projects/glove/ which were trained on the Common Crawl (840B tokens) dataset. The authors in [51] trained GloVe on the 2016 PubMed baseline and made them publicly available at https://slate.cse.ohio-state.edu/BMASS/.

FastText General embeddings were obtained from https://fasttext.cc/docs/en/english-vectors.html also trained on the Common Crawl corpus resulting in 2 million word vectors. For the domain-specific pre-trained model, we used the embeddings provided at https://github.com/lucylw/pubmed_central_fasttext_pretrained.

ELMo We use the original 5.5B configuration, as recommended by the authors, trained on Wikipedia and news crawl data. To further investigate the embedding size effect on performance, we added the small model trained on the 1 Billion Word Benchmark. Moreover, we download their biomedical domain contributed model trained on PubMed. ELMo embeddings are computed after concatenating all three layers of the ELMo. All models were downloaded from https://allennlp.org/ELMo and implemented through the AllenNLP python toolkit [52].

BERT We evaluated both base and large base models provided at https://github.com/google-research/bert. We also take advantage of two newly released BERT models: BioBERT [53] trained on the PubMed abstracts with a vocabulary size of 4.5B words and SciBERT [54] trained on scientific articles from the biomedical and computer sciences domains with 2.5B and 0.6B word count, respectively. The pre-trained weights of the BioBERT model (version 1.0/ PubMed 200 K) are available at https://github.com/naver/biobert-pretrained and for the SciBERT model at https://github.com/allenai/scibert. We use the original Google BERT GitHub repository to encode sentences; it originally provides fine-tuning scripts for the pre-trained model in an end-to-end fashion. It additionally describes how to obtain fixed contextual embeddings of each input token generated from the hidden layers of the pre-trained model. Following the steps to obtain the embeddings, we only use the final hidden layer of the transformer (layer value set to −1). The maximum sequence length was set to 128 with a batch size of 32 as per the authors' recommendation.

Flair The authors recommend using both forward and backward Flair embeddings. We chose the mixed model as it is trained over a diverse corpus including web, Wikipedia, and Subtitles for the English language. They also provide pre-trained embeddings over 5% of PubMed abstracts until 2015. All models are downloaded from and implemented through the official Flair repository https://github.com/zalandoresearch/flair.

InferSent The authors provide two versions of the model, one based on GloVe embeddings and another based on FastText embeddings. We experiment with both available at https://github.com/facebookresearch/InferSent. InferSent is based on supervised learning from natural inference data. For this model, we train our own models using the Medical natural language inference (MedNLI) dataset using the biomedical Glove and FastText embeddings mentioned earlier.

USE In our evaluation, we use the transformer-based architecture of the USE encoder as it was proven to yield better results. Training data consisted of supervised and unsupervised sources such as Wikipedia articles, news, discussion forums, dialogues and question/answer pairs. USE was implemented through its TF hub module available at https://tfhub.dev/google/universal-sentence-encoder-large/3.

3.3. Experimental setup

A fundamental dilemma is how to compare different models that vary between vanilla embeddings, contextualized embeddings, or dedicated sentence encoders. The performance differences between models could be attributed to many reasons such as better pre-trained word embeddings, the different architecture, the different objective, or the normalization layer [55]. SentEval was created to overcome the limitation of the non-standard evaluation of embedding methods across research, especially when a considerable number of embeddings techniques have surfaced in the last years. SentEval's evaluation strategy relied on simple classification models. The choice behind the basic classifiers was to assess how these representations fare without the use of complex models like recurrent neural networks (RNNs). As in the original SentEval tooklit, we opted for the same models as it allows us to observe the benefits of different embedding models in representing and predicting linguistic medical information of the input text. Our choice was also supported by several studies following the SentEval methodology in evaluating new models [8,55–57]. We follow the same guidelines in our experimental settings. For tasks that require classification, we experiment with both logistic regression and multi-layer perceptron (MLP) on top of the generated sentence representations. The MLP consists of a single hidden layer of 50 neurons using Adam optimizer and a batch size of 64 with a Sigmoid non-linearity function. Extracting word embeddings from vanilla models such as GloVe and FastText is a straightforward process. On the other hand, the contextualized models offer two paradigms to deploy their pre-trained models to adapt to the target task: feature extraction and fine-tuning. The former is similar to feature-based models where pre-trained weights are kept frozen whereas, in the latter, the weights are trained further on the new task. In a recent paper, Peters et al. compare the effectiveness of both adaptation methods. Their results show that BERT generally performed better in the feature-extraction mode while the opposite is exact for ELMo. Their evaluation also proves that the performance depends on the similarity of the pre-training and target tasks. The feature-extraction mode aligns with our strategy to standardize the comparison criteria across all evaluation experiments. Not only because the use of embeddings as features is the only possible method for other models but also because the fine-tuning process differs from BERT to ELMo. In MedSentEval, ELMo features are calculated for each token by concatenating all three layers weights of the model. For BERT features, we take the hidden states of the final hidden layer of the transformer model. To generate sentence vectors from word embeddings, we apply the Mean of Word Embeddings (MOWE) technique [58] on both static and contextualized embeddings. In tasks with dual inputs as in textual entailment tasks, their combined embedding vector is built as $(u, v, |u - v|, u * v)$, which is a concatenation of the premise and hypothesis vectors and their respective absolute difference and Hadamard product. Tasks evaluation criteria are consistent across tasks with minor variations as their input/output formats, and types are different. For example, for semantic similarity tasks, we only need to calculate the cosine similarity between the input's embedding and compare it to the expert-labeled score through Pearson and Spearman correlations. All experiments were carried out using a single GPU with 12 GB RAM. However the toolkit also provides optional support of CPU only machines through scikit-learn for the logistic regression. Since this might trigger memory issues with some datasets such as *PICO* and *PubMed20K*, we only recommend this for small sized datasets. Table 2 highlights the experimental settings for each task and the performance metrics used for evaluation. Our analysis do not include the time factor when conducting the comparison since both the feature extraction and classification phases do not exceed 2 h for all tasks with the exception of BERT and ELMo models when run over big datasets namely *PICO* and *PubMed20K*. We note that this does not include the time needed to generate the pre-trained weights as it may cost more time to train some embedding models, such as InferSent, Bert or ELMo. (see Fig. 1).

4. Results

In Table 3, shows results obtained from the included embedding schemes across all 10 tasks included in *MedSentEval*. The reported results are based on the logistic regression classifier as it consistently achieved better results than MLP on most transfer tasks and specifically on small size datasets. ELMo takes the lead with the original 5.5B model excelling in 5 out of 10 tasks in the general embeddings category. The PubMed version is also dominating with 4 tasks in the embeddings acquired from biomedical training data. The BERT algorithm comes next with the best performance of 2 and 3 tasks for base and BioBert models respectively. Moreover, BERT embeddings are often the second best performing on many tasks with minimal accuracy difference from ELMo which did not exceed 1%. Fig. 2 and 3 illustrate a comparison between each method with general and domain-specific training. Additionally, we investigate whether there is any correlation between independent model factors and perceived performance. Driven by several research questions, we analyzed the results of the conducted evaluation.

Static versus Context embeddings Comparing the representation models within each category separately, we find that context-dependent models capture more information than regular static embeddings. The only exception to that rule was the BioASQ dataset, where Glove and FastText achieve better results than ELMo, BERT and Flair in the biomedical domain and BERT and ELMo only in the general domain. This exception is not indicative as we additionally observe that tasks in the question-answering category, in general, are the least influenced by the different techniques since the classifier tends to overfit to the majority class in most models.

General versus Domain embeddings Apart from GloVe, FastText, and Flair, comparing each general embedding model to its biomedical peer, the latter always outperforms the former. In the case of Flair, the medical embeddings are worst in performance or do not provide a significant gain over the general model. As mentioned, ELMo and BERT are the best-suited models in both the general and biomedical categories to represent medical text.

Word-based versus Sentence-encoders embeddings Under the assumption that sentence-level encodings better capture the content of medical text since it takes into account the word order within the sentence, we observed that the best results are generally obtained through averaging word embeddings. The reason for this result may be related to the fact that much of the word order information is captured in general natural language word order statistics [23]. This observation is true for all tasks except for the language inference task. However, we believe that this is due to the similarity of the task's data and the training data of the sentence encoders. The inferSent supervised model is trained on the SNLI and MedNLI datasets for the general and biomedical embeddings categories, respectively. While the Universal Sentence Encoder has multi-type data including questions and entailment pairs, among others.

Embedding Dimension versus Embedding model As we employ the averaging scheme to calculate the sentence embedding, the size of the embedding vector is equal to the original word vector size. While FLAIR and InferSent have the biggest embedding dimension of 4096, their performance is inferior to other models with a much smaller embedding vector. On the other hand, in models like ELMo and BERT, where we experiment with different versions of the same model with different embedding dimensions, we notice that increasing the embedding vector size is related to a performance gain. This finding is expected as the more dimensions a word vector has, the more semantic information can be preserved in the resulting sentence representation. This also might explain the poor performance of GloVe and FastText biomedical embeddings as opposed to their corresponding general models. The embedding size of the pre-trained biomedical model are 200 and 100 for GloVe, and FastText respectively, while the general embedding dimension is 300.

Table 2
Experimental settings for each evaluation task.

Task	Classes	Classification	Validation	Performance Metrics
MedNLI	3	LR/MLP	Standard validation	Accuracy
RQE	2	LR/MLP	Standard validation	Accuracy
PUBMED20K	5	LR/MLP	Standard validation	Accuracy
PICO	8	LR/MLP	Standard validation	Accuracy
PatientSA	8	LR/MLP	Nested cross-validation	Accuracy
CitationSA	3	LR/MLP	Nested cross-validation	Accuracy
BioASQ	2	LR/MLP	Cross-validation	Accuracy/F1
BioC	2	LR/MLP	Nested cross-validation	Accuracy/F1
C-STS	[0–5]	–	Cosine similarity	Pearson/Spearman correlation
BIOSSES	[0–4]	–	Cosine similarity	Pearson/Spearman correlation

4.1. Qualitative analysis

Besides the quantitative results mentioned above, we also test the effectiveness of the different biomedical models on several simple, but non-trivial, examples beyond the extrinsic tasks. In this section, we shed light on intrinsic characteristics that may explain some of the performance variances. Building on our previous findings that domain-based embeddings are better suited for the medical and clinical NLP tasks, we limit the qualitative analysis to the biomedical embedding models.

The diagnostic data used throughout this analysis have been manually selected to fit the test purpose. However, they do not represent the language distribution as a whole. Our main goal was to provide insight into what models are capturing, what are their strengths and weaknesses through adversarial examples.

General Knowledge We first attempt to investigate the robustness of the generated numerical representations to reflect common sense [59]. The test consists of removing non-important tokens, such as stop words, and calculate the similarity between the original and the shortened sentences. For this test, we collect ten sentences, one from each dataset, while varying the number of stop words per chosen sentences. Table 4 shows an example of the sentences before and after removal, and the corresponding cosine distances computed by the six biomedical models. The models are ranked according to the descending order of the similarity values. The higher the value, the better the ability of the model to assign similar embeddings for both sentences. Consider the example in Table 4, most models still retains a similarity of 0.93 or above, except for GloVe, even after removing 13 stop words from a single sentence. More examples are available in Appendix A. The results demonstrate that the InferSent model is the most insensitive regarding the removal of non-important words and, surprisingly, ELMo's performance is not consistent. In many cases, the similarity decreases by 10%, ranking below FastText in all 10 examples. The same applies to Flair embeddings; this suggests that preprocessing the text before embedding with these models could give higher priority to informative words and might yields better results in downstream tasks.

Concept Identity The second test measures to what extent the sentence representation encodes the identities of entities within it. In the medical and clinical language, referring to concepts using their abbreviations is a common practice and frequently found in sentences extracted from both patient records and scientific articles. Retaining the concept identity, whether it is referred to in full or in abbreviated form, is crucial and demonstrates the models' capacities pertinent to language understanding. We collected five examples with abbreviated concepts in the premises from the MedNLI dataset. We compared the NLI predicted labels given the original premise and after expanding the abbreviations. It is clear from observing the examples in Table 4 and Appendix A that BERT is able to relate the premise to the hypothesis more often when using the full-form leading to the correct inference. On the other hand, ELMo has zero gain after the expansion process. While the number of cases is relatively small to generalize the results, based on this intuition, we suggest expanding and normalizing abbreviations and acronyms in the dataset before using BERT.

Domain Knowledge The third test assesses the model's ability to capture significant semantic meanings of the input text. In the context of our domain-based evaluation, the semantic significance mainly refers to the medical information within the sentence [60]. For this test, we consider 6 records that hold indirect information relating independent medical concepts. In the example shown in Table 4, most models fail to relate *chest pain* to *Angina*. Consequently they interpret these to be two unrelated entities and labels the instance pair as neutral. This pattern is consistent in most models across different relation categories such as *Disease-Symptom*, *Drug-Disease*, *Drug-Drug classes*. While empirical results show that all the representations encode certain amount of information, our finding advocates for integrating external knowledge to compensate for the lack of medical background of the models. This could be achieved by adding external knowledge sources such as the Unified Medical Language System (UMLS) to include semantic types and relationships.

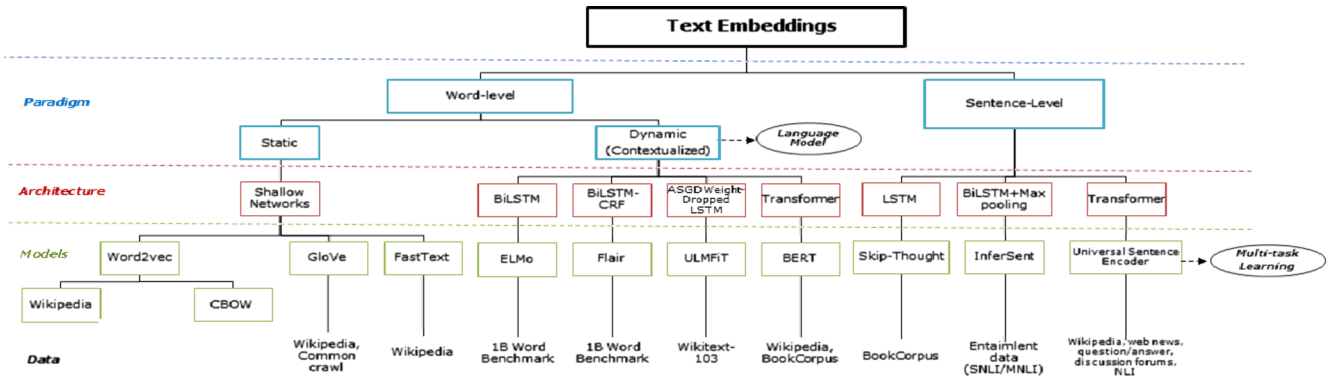


Fig. 1. Classification of different embedding models.

Table 3
Logistic regression performance on all tasks included in *MedSentEval*. For all tasks we report accuracies except for the BioC/BioASQ we additionally report F1. And for BIOSSES and ClincalSTS, we report Pearson/Spearman correlations between the cosine distance of both sentences and the similarity score given by the domain expert. Underlined values indicate the best overall result, while values in **bold** indicate best performing over each category.

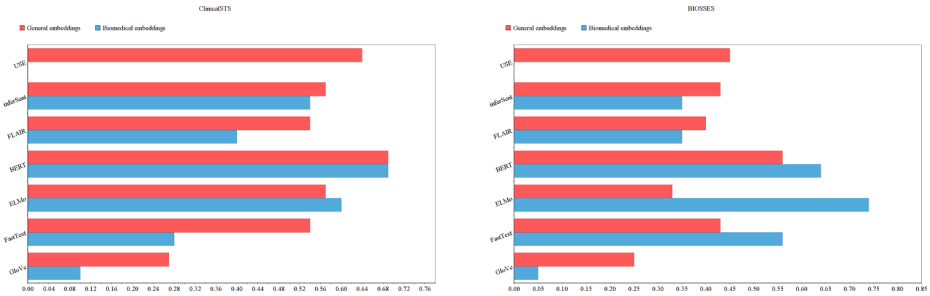
Tasks	Emb. size	MedNLI	RQE	PubMed20K	PICO	Vac.SA	Cit.SA	BioC	BioASQ	C-STS	BIOSSES
General Embeddings											
$GloVE_{840B}$	300	65.05	60.93	73.95	73.95	60.20	78.97	69.89/82.20	69.23/81.82	0.27/0.48	0.25/0.39
$FastText_{crawl}$	300	63.08	62.25	76.23	75.09	61.67	79.39	69.51/82.01	69.23/81.82	0.54/0.56	0.43/0.50
$ELMo_{small}$	768	62.03	58.28	80.26	78.68	62.22	81.73	74.12/83.48	66.15/78.35	0.54/0.50	0.33/0.35
$ELMo_{Org5.5B}$	3072	66.10	57.92	**83.51**	**81.89**	**68.82**	**83.60**	**81.44/87.81**	67.02/78.80	0.57/0.49	0.29/0.32
$Bert_{baseCased}$	768	63.99	65.56	81.91	80.73	64.54	83.18	74.89/84.66	66.92/79.62	**0.69/0.57**	0.48/0.50
$Bert_{largeUncased}$	1024	65.54	68.21	83.04	81.08	66.07	82.80	74.51/84.09	69.23/81.30	0.66/0.51	**0.56/0.58**
$Flair_{news}$	4096	61.12	56.62	81.58	80.15	67.21	82.79	72.20/83.30	66.92/79.62	0.54/0.47	0.31/0.30
$Flair_{mix}$	4096	64.42	58.94	81.26	79.60	65.62	81.54	70.28/82.39	**70.00/82.03**	0.52/0.53	0.40/0.50
$InferSent_1$	4096	**67.16**	58.22	74.70	61.42	66.91	81.86	79.94/97.18	70.00/81.86	0.57/0.54	0.32/0.42
$InferSent_2$	4096	63.99	62.25	78.24	78.24	64.28	80.55	69.51/82.01	68.46/81.28	0.54/0.57	0.43/0.48
USE	512	60.76	**73.84**	75.50	73.26	62.46	78.76	69.50/82.01	69.23/81.82	0.64/0.56	0.45/0.48
Biomedical embeddings											
$Glove_{PubMed}$	200	56.96	56.29	66.61	65.26	45.08	78.71	69.50/82.01	**69.23/81.82**	0.08/0.42	0.05/0.17
$FastText_{PubMed}$	100	61.39	63.25	75.47	74.16	55.67	78.92	69.50/82.01	**69.23/81.82**	0.28/0.56	0.56/0.60
$ELMo_{PubMed}$	3072	71.18	62.91	**85.71**	83.76	**68.79**	84.73	**83.00/88.34**	66.92/80.00	0.61/0.54	**0.74/0.70**
BioBERT	768	66.95	63.91	85.35	83.69	65.50	83.91	74.51/84.21	68.46/81.28	**0.69/0.54**	0.64/0.62
SciBERT	768	66.24	63.91	85.44	**83.77**	65.86	**84.93**	82.25/88.41	67.69/80.00	0.67/0.59	0.56/0.60
$Flair_{PubMed}$	2300	61.60	57.95	82.73	81.08	57.61	81.81	72.61/83.47	66.92/79.81	0.40/0.48	0.35/0.47
$InferSent_{1MedNLI}$	4096	**71.52**	**66.23**	79.54	78.86	63.47	79.80	71.05/82.69	68.85/81.55	0.54/0.53	0.35/0.41

(a) MEDNLI (b) RQE (c) PubMed20K

(d) PICO (e) VaccineSA (f) CitationSA

(g) BioC (h) BioASQ

Fig. 2. Accuracy values for the logistic regression classifier across tasks included in *MedSentEval*.



(a) C-STS (b) BIOSSES

Fig. 3. Results for the Semantic Text Similarity tasks. Values shown are the Pearson correlation coefficients for the test sets.

5. Discussion

This paper inspects sentence representations for biomedical text by analyzing seven popular embedding schemes. It is important to recall that the primary purpose of this study is not to outperform existing state-of-the-art methods for the reported BioNLP tasks. We are seeking to evaluate and validate different embedding techniques that will enable further in-depth investigations and improvement of text representations for the biomedical domain. When comparing with baseline performance introduced in datasets' original papers, if applicable, the results obtained is close to or outperforms baseline values. This is mainly due to the use of simple classification algorithms like the MLP or LR classifiers. Therefore, the performance achieved through the toolkit still has room for improvement by fine tuning the hyperparameters of the classification models or deploying other classifiers. It was reported that employing complex models such as Convolutional Neural Networks (CNNs) is more effective for text classification tasks [61]. Similarly, using the end-to-end BERT model could lead to a non-trivial accuracy gain for several tasks such as MedNLI [49,63]. However, the effect of classification algorithms on the performance and the analysis of different approaches to adapt the pre-trained representations [64] are out of the scope of this paper.

The benchmark gives insights on the sentence embedding quality through downstream tasks with four extrinsic tasks (Textual Entailment, Sentence classification, Sentiment analysis, and Question answering). Additionally, we follow a case-based reasoning approach to provide a qualitative analysis of the learned representations.

We found relatively modest correlations between the quantitative and qualitative results. In the case of ELMo, for example, the intrinsic evaluation results fail extrinsic performance. This also matches previous evaluations outside the medical domain [65]. Finally, both evaluation types could benefit from domain experts' perspectives. It is hard to draw conclusions or rank models from best to worst as there is no single sentence embedding scheme that consistently performs well on all of the ten tasks. As with all classification problems, specific approaches are better suited to some datasets than others, this is also consistent with the "no free lunch theorem" [66]. However, our experimental results unveil a number of important observations:

- Sentence embeddings computed as the mean of word embeddings are still effective in capturing the sentence semantics and yield competitive results to dedicated sentence encoders.
- There is no correlation between the embedding dimension and the performance across different models.
- In almost all cases, neural embeddings generated from hidden states of a deep learning model are able to capture more semantics than word embeddings computed from count or prediction based models.
- Contextualized word embeddings with a language model objective, i.e. ELMo and BERT, usually outperform other encoding schemes.
- While InferSent is better suited for textual entailment, given the type of data it is trained on, its good performance does not generalize over other tasks.
- A proper balance and variation in the training resources, when compared to training solely on domain data, can lead to more efficient results such as the case of BioBert and SciBERT.

The results show that most models still need to resume training on domain-related and task-specific data. And that, to date, producing a single universal embedding model that generalizes well to other tasks requires more investigations and evaluations. Given the superiority of

Table 4
Qualitative Analysis.

Test Objective	Example	Predicted			Expected
General Knowledge	*Original: Our findings suggest an association between the DD genotype of the ACE gene and early-onset but not later-onset pre-eclampsia which may give a partial explanation for the higher recurrence risk with early- onset pre-eclampsia.* *Modified: Our findings suggest association DD genot-ype ACE gene early-onset later-onset pre-eclampsia may give partial explanation higher recurrence risk early- onset pre-eclampsia.*	FastText InferSent BERT ElMo Flair Glove	0.97 0.97 0.96 0.87 0.84 0.8		~1
Concept Identity	*Premise: Reports lack of appetite but no n/v.* *Expanded premise: Reports lack of appetite but no nausea and vomiting.* *Hypothesis: the patient denies nausea and vomiting.*		Pre	Post	E
		GloVe	N	C	
		FastText	N	E	
		ElMo	C	C	
		BERT	C	E	
		Flair	N	C	
		InferSent	C	E	
Medical Knowledge	*Premise: No chestpain or fevers.* *Hypothesis: Patient has no angina*	GloVe FastText ElMo BERT Flair InferSent	C E N N C E		E

ELMo and BERT over other models, we particularly recommend integrating language models with neural embeddings as a promising direction of research. Incorporating medical knowledge in the learning process of the models is also expected to enhance their performance. Another alternative for improving the results is to combine two or more embedding techniques in a single classification model. Adopting such a method could offer specific domain background, when adding concept embeddings for example, to acquire the best of each technique.

6. Conclusion

In this paper, we presented *MedSentEval*, a new toolkit for evaluating state-of-the-art sentence embedding methods for NLP classification problems. Through our evaluations, we assessed the transferability of these embeddings to biomedical domain tasks. Our research aimed to build on the work done by Conneau et al. [7] and adapt it to fit medical and clinical text corpora. We also integrated extra embedding techniques not available in the original toolkit such as ELMo and BERT. We hope that our in-depth evaluations, along with the toolkit, will benefit the BioNLP community in selecting suitable embeddings for different application tasks. While only downstream tasks are used to evaluate the overall quality of sentence representation models, we also note the need to incorporate probing tasks as in *SentEval*. A future extension of our current work will include support for more embeddings schemes trained on different domain data types such as patient records, nurse notes and full-text articles PubMed central combined. Moreover, adding more tasks for each category, when available, could further improve our understanding and generalization of the findings.

CRediT authorship contribution statement

Noha S. Tawfik: Conceptualization, Data curation, Investigation, Methodology, Software, Writing - original draft. **Marco R. Spruit:** Project administration, Supervision, Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

xxxx

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jbi.2020.103396.

References

[1] T. Schnabel, I. Labutov, D. Mimno, T. Joachims, Evaluation methods for unsupervised word embeddings, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 298–307, , https://doi.org/10.18653/v1/D15-1036 https://www.aclweb.org/anthology/D15-1036.

[2] Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, P. Kingsbury, H. Liu, A comparison of word embeddings for the biomedical natural language processing, J. Biomed. Inform. 87 (2018) 12–20, https://doi.org/10.1016/J.JBI.2018.09.008 https://www.sciencedirect.com/science/article/pii/S1532046418301825?via%3Dihub.

[3] Z. Chen, Z. He, X. Liu, J. Bian, Evaluating semantic relations in neural word embeddings with biomedical and general domain knowledge bases, BMC Med. Inform. Decis. Mak. 18 (S2) (2018) 65, https://doi.org/10.1186/s12911-018-0630-x https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-018-0630-x.

[4] B. Chiu, G. Crichton, A. Korhonen, S. Pyysalo, How to Train Good Word Embeddings for Biomedical NLP, in: Proceedings of the 15th Workshop on Biomedical Natural Language Processing, Berlin, Germany, 2016, pp. 166–174. doi:10.18653/v1/W16-2922. https://www.aclweb.org/anthology/W16-2922.

[5] Q. Chen, Y. Peng, Z. Lu, BioSentVec: creating sentence embeddings for biomedical texts, arXiv e-prints. http://arxiv.org/abs/1810.09302.

[6] Y. Hao, X. Liu, J. Wu, P. Lv, Exploiting Sentence Embedding for Medical Question Answering, arXiv e-prints. http://arxiv.org/abs/1811.06156.

[7] A. Conneau, D. Kiela, SentEval: An evaluation toolkit for universal sentence representations, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018). https://aclanthology.info/papers/L18-1269/l18-1269.

[8] C.S. Perone, R. Silveira, T.S. Paula, Evaluation of sentence embeddings in downstream and linguistic probing tasks, arXiv e-prints. https://arxiv.org/pdf/1806.06259.pdf.

[9] R. Mackin, On collocations: words shall be known by the company they keep, in: P. Strevens (Ed.), In honor of A.S. Hornby, Oxford University Press, London, Oxford, 1978, pp. 149–165.

[10] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, J.U. Ca, J. Kandola, T. Hofmann, T. Poggio, J. Shawe-Taylor, A neural probabilistic language model, J. Mach. Learn. Res. 3 (2003) 1137–1155 http://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf.

[11] T. Mikolov, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Adv. Neural Inf. Process. Syst. (2013) 3111–3119.

[12] Jeffrey Pennington, Richard Socher, Christopher D Manning, GloVe: global vectors for word representation, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014) 1532–1543, https://doi.org/10.3115/v1/D14-1162 https://www.aclweb.org/anthology/D14-1162.

[13] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Trans. Assoc. Comput. Linguist. 5 (2017) 135–146, https://doi.org/10.1162/tacl_a_00051 https://www.aclweb.org/anthology/Q17-1010.

[14] Y. Yaghoobzadeh, H. Schütze, Intrinsic Subspace Evaluation of Word Embedding Representations, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Stroudsburg, PA, USA, 2016, pp. 236–246. doi:10.18653/v1/P16-1023. URL http://aclweb.org/anthology/P16-1023.

[15] O. Melamud, J. Goldberger, I. Dagan, context2vec: Learning Generic Context Embedding with Bidirectional LSTM, in: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, Association for Computational Linguistics, Stroudsburg, PA, USA, 2016, pp. 51–61. doi:10.18653/v1/K16-1006. http://aclweb.org/anthology/K16-1006.

[16] M.E. Peters, M. Neumann, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2227–2237. https://www.aclweb.org/anthology/N18-1202.

[17] J. Devlin, M.-W. Chang, K. Lee, K.T. Google, A.I. Language, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv e-prints. https://github.com/tensorflow/tensor2tensor.

[18] J. Howard, S. Ruder, Universal Language Model Fine-tuning for Text Classification, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Association for Computational, Linguistics, Melbourne, 2018, pp. 328–339 https://www.aclweb.org/anthology/P18-1031.

[19] A. Akbik, D. Blythe, R. Vollgraf, Contextual String Embeddings for Sequence Labeling, Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 1638–1649.

[20] M. Pagliardini, P. Gupta, M. Jaggi, Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Stroudsburg, PA, USA, 2018, pp. 528–540. doi:10.18653/v1/N18-1049. http://aclweb.org/anthology/N18-1049.

[21] S. Arora, Y. Liang, T. Ma, A Simple but Tough-to-Beat Baseline for Sentence Embeddings, Proceedings of the International Conference on Learning Representations, 2017.

[22] H. Li, D. Caragea, X. Li, C. Caragea, Comparison of Word Embeddings and Sentence Encodings as Generalized Representations for Crisis Tweet Classification Tasks, Proceedings of the ISCRAM Asian Pacific 2018 Conference, 2018.

[23] Y. Adi, E. Kermany, Y. Belinkov, O. Lavi, Y. Goldberg, Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks, 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings, 2017.

[24] R. Kiros, Y. Zhu, R. Salakhutdinov, R.S. Zemel, A. Torralba, R. Urtasun, S. Fidler, Skip-Thought Vectors, in: Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, Canada, 2015, pp. 3294–3302. https://papers.nips.cc/paper/5950-skip-thought-vectors.pdf.

[25] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, A. Bordes, Supervised learning of universal sentence representations from natural language inference data, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 2017.

[26] D. Cer, Y. Yang, S.-Y. Kong, N. Hua, N. Limtiaco, R. St John, N. Constant, M. Guajardo-Céspedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, R. Kurzweil Google research mountain view, universal sentence encoder, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Brussels, 2018, pp.

169–174. https://www.aclweb.org/anthology/D18-2029.
[27] I. Dagan, D. Roth, M. Sammons, F.M. Zanzotto, Recognizing textual entailment: models and applications, Synthesis Lect. Hum. Lang. Technol. 6 (4) (2013) 1–220, https://doi.org/10.2200/S00509ED1V01Y201305HLT023 http://www.morganclaypool.com/doi/abs/10.2200/S00509ED1V01Y201305HLT023.
[28] A. Romanov, C. Shivade, Lessons from natural language inference in the clinical domain, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 1586–1596.
[29] A Natural Language Inference Dataset For The Clinical Domain. (Accessed 04 March 2019).
[30] A.E. Johnson, T.J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Scientific Data 3. doi:10.1038/sdata.2016.35. http://www.nature.com/articles/sdata201635.
[31] A. Ben Abacha, D. Demner-Fushman, Recognizing question entailment for medical question answering, AMIA Annual Symposium Proceedings 2016, 2016, pp. 310–318 http://www.ncbi.nlm.nih.gov/pubmed/28269825.
[32] The Medical Question Entailment Data. (Accessed 03 March 2019).
[33] D. Jin, P. Szolovits, PICO Element Detection in Medical Text via Long Short-Term Memory Neural Networks, Tech. rep., 2018. http://www.aclweb.org/anthology/W18-2308.
[34] D. Jin, P. Szolovits, Pico Element Detection in Medical Text via Long Short-term Memory Neural Networks, Proceedings of the BioNLP 2018 workshop, 2018, pp. 67–75.
[35] PubMed PICO Element Detection Dataset. (Accessed 05 March 2019).
[36] PubMed 200k RCT Dataset. (Accessed 03 March 2019).
[37] Franck Dernoncourt, Ji Young Lee, PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts, Proceedings of the Eighth International Joint Conference on Natural Language Processing (2017) 308–313 https://www.aclweb.org/anthology/I17-2052.
[38] J. Xu, Y. Zhang, Y. Wu, J. Wang, X. Dong, H. Xu, Citation sentiment analysis in clinical trial papers, AMIA Annual Symposium proceedings. AMIA Symposium 2015, 2015, pp. 1334–1341 http://www.ncbi.nlm.nih.gov/pubmed/26958274.
[39] Citation Sentiment Analysis Dataset (personal communication). (Accessed 28 February 2019).
[40] J. Du, J. Xu, H. Song, X. Liu, C. Tao, Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets, J. Biomed. Semantics 8 (1) (2017) 9, https://doi.org/10.1186/s13326-017-0120-6 http://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-017-0120-6.
[41] HPV Vaccination's Tweets Dataset. (Accessed 09 March 2019).
[42] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M.R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, Y. Almirantis, J. Pavlopoulos, N. Baskiotis, P. Gallinari, T. Artiéres, A.-C.N. Ngomo, N. Heino, E. Gaussier, L. Barrio-Alvers, M. Schroeder, I. Androutsopoulos, G. Paliouras, An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition, BMC Bioinformatics 16 (1) (2015) 138, https://doi.org/10.1186/s12859-015-0564-6 https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0564-6.
[43] Biomedical Semantic Question Answering Dataset. (Accessed 03 March 2019).
[44] A. Alamri, M. Stevenson, A corpus of potentially contradictory research claims from cardiovascular research abstracts, J. Biomed. Semantics 7 (36) (2016), https://doi.org/10.1186/s13326-016-0083-z http://www.ncbi.nlm.nih.gov/pubmed/27267226 , http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4897929.
[45] A Corpus of Contradictory Research Claims from Cardiovascular Research Abstracts. (Accessed 11 March 2019).
[46] Y. Wang, S. Liu, M. Rastegar-Mojarad, N. Afzal, L. Wang, F. Shen, S. Fu, H. Liu, Overview of BioCreative/OHNLP Challenge 2018 Task 2: Clinical Semantic Textual Similarity, in: Proceedings of the BioCreative/OHNLP Challeng, Washington, 2018. doi:10.13140/RG.2.2.26682.24006. https://github.com/ohnlp/BioCreativeOHNLPProceedings/raw/master/clinicalsts_overview.pdf.
[47] Clinical Semantic Textual Similarity Dataset (Retrieved through personal communication). (Accessed 03 February 2019).
[48] Y. Wang, N. Afzal, S. Fu, L. Wang, F. Shen, M. Rastegar-Mojarad, H. Liu, MedSTS: a resource for clinical semantic textual similarity, Lang. Resources Eval. (2018) 1–16, https://doi.org/10.1007/s10579-018-9431-1 http://link.springer.com/10.1007/s10579-018-9431-1.
[49] G. Soganciogl u, H. Öztürk, A. Özgür, BIOSSES: a semantic sentence similarity estimation system for the biomedical domain, Bioinformatics (Oxford, England) 33 (14) (2017) i49–i58, https://doi.org/10.1093/bioinformatics/btx238 http://www.ncbi.nlm.nih.gov/pubmed/28881973.
[50] Biomedical Semantic Similarity Estimation System. (Accessed 28 February 2019).
[51] D. Newman-Griffis, A.M. Lai, E. Fosler-Lussier, Insights into analogy completion from the biomedical domain, Proceedings of the 16th Workshop on Biomedical Natural Language Processing (BioNLP), 2017, pp. 19–28.
[52] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N.F. Liu, M. Peters, M. Schmitz, L. Zettlemoyer, AllenNLP: A Deep Semantic Natural Language Processing Platform, in: Proceedings of Workshop for NLP Open Source Software (NLP-OSS), Association for Computational Linguistics, Melbourne, 2018, pp. 1–6. https://aclweb.org/anthology/papers/W/W18/W18-2501/.
[53] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, arXiv e-prints. http://arxiv.org/abs/1901.08746.
[54] I. Beltagy, A. Cohan, K. Lo, SCIBERT: Pretrained Contextualized Embeddings for Scientific Text, arXiv e-prints. https://arxiv.org/abs/1903.10676.
[55] J. Wieting, D. Kiela, No Training Required: Exploring Random Encoders for Sentence Classification, International Conference on Learning Representations, 2019.
[56] N. Reimers, I. Gurevych, Sentence-BERT: sentence embeddings using siamese BERT-networks, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019.
[57] J. Kiros, W. Chan, InferLite: simple universal sentence representations from natural language inference data, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, 2018, pp. 4868–4874. doi:10.18653/v1/D18-1524. http://aclweb.org/anthology/D18-1524.
[58] L. White, R. Togneri, W. Liu, M. Bennamoun, M. Ben, How well sentence embeddings capture meaning, in: Proceedings of the 20th Australasian Document Computing, ACM, Parramatta, NSW, Australia, 2015. doi:10.1145/2838931.2838932.
[59] Z. Yang, C. Zhu, W. Chen, Parameter-free sentence embedding via orthogonal basis, Assoc. Comput. Linguist. (ACL) (2019) 638–648, https://doi.org/10.18653/v1/d19-1059.
[60] W.-H. Weng, P. Szolovits, Representation Learning for Electronic Health Records. http://arxiv.org/abs/1909.09248.
[61] Y. Kim, Convolutional Neural Networks for Sentence Classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Stroudsburg, PA, USA, 2014, pp. 1746–1751. doi:10.3115/v1/D14-1181. http://aclweb.org/anthology/D14-1181.
[62] E. Alsentzer, J.R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, M.B.A. McDermott, Publicly Available Clinical BERT Embeddings, arXiv e-prints (2019) arXiv:1904.03323. http://arxiv.org/abs/1904.03323.
[63] N. Tawfik, M. Spruit, UU_TAILS at MEDIQA 2019: Learning Textual Entailment in the Medical Domain, in: Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, Florence, Italy, 2019. doi:10.18653/v1/W19-5053. https://www.aclweb.org/anthology/W19-5053.
[64] M. Peters, S. Ruder, N.A. Smith, To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks, arXiv e-prints. https://arxiv.org/pdf/1903.05987.pdf.
[65] N. Hollenstein, A. de la Torre, N. Langer, C. Zhang, CogniVal: a framework for cognitive word embedding evaluation, Assoc. Comput. Linguist. (ACL) (2019) 538–549, https://doi.org/10.18653/v1/k19-1050.
[66] D.H. Wolpert, W.G. Macready, No free lunch theorems for optimization, IEEE Trans. Evol. Comput. 1 (1) (1997) 67 https://ti.arc.nasa.gov/m/profile/dhw/papers/78.pdf.