frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 12 August 2022
DOI 10.3389/fmolb.2022.928530

Medical terminology-based computing system: a lightweight post-processing solution for out-of-vocabulary multi-word terms

Nadia Saeed and Hammad Naveed*

Computational Biology Research Lab, Department of Computer Science, National University of Computer and Emerging Sciences (NUCES-FAST), Islamabad, Pakistan

OPEN ACCESS

EDITED BY
Feng Gao,
The Sixth Affiliated Hospital of Sun Yat-sen University, China

REVIEWED BY
Shuting Yang,
Second Xiangya Hospital, Central South University, China
Zeliha Selamoglu,
Niğde Ömer Halisdemir University, Turkey

*CORRESPONDENCE
Hammad Naveed,
hammad.naveed@nu.edu.pk

SPECIALTY SECTION
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 25 April 2022
ACCEPTED 04 July 2022
PUBLISHED 12 August 2022

CITATION
Saeed N and Naveed H (2022), Medical terminology-based computing system: a lightweight post-processing solution for out-of-vocabulary multi-word terms.
Front. Mol. Biosci. 9:928530.
doi: 10.3389/fmolb.2022.928530

The linguistic rules of medical terminology assist in gaining acquaintance with rare/complex clinical and biomedical terms. The medical language follows a Greek and Latin-inspired nomenclature. This nomenclature aids the stakeholders in simplifying the medical terms and gaining semantic familiarity. However, natural language processing models misrepresent rare and complex biomedical words. In this study, we present MedTCS—a lightweight, post-processing module—to simplify hybridized or compound terms into regular words using medical nomenclature. MedTCS enabled the word-based embedding models to achieve 100% coverage and enabled the BiowordVec model to achieve high correlation scores (0.641 and 0.603 in UMNSRS similarity and relatedness datasets, respectively) that significantly surpass the n-gram and sub-word approaches of FastText and BERT. In the downstream task of named entity recognition (NER), MedTCS enabled the latest clinical embedding model of FastText-OA-All-300d to improve the F1-score from 0.45 to 0.80 on the BC5CDR corpus and from 0.59 to 0.81 on the NCBI-Disease corpus, respectively. Similarly, in the drug indication classification task, our model was able to increase the coverage by 9% and the F1-score by 1%. Our results indicate that incorporating a medical terminology-based module provides distinctive contextual clues to enhance vocabulary as a post-processing step on pre-trained embeddings. We demonstrate that the proposed module enables the word embedding models to generate vectors of out-of-vocabulary words effectively. We expect that our study can be a stepping stone for the use of biomedical knowledge-driven resources in NLP.

KEYWORDS

medical terminology, named entity recognition, linguistic approach, natural language processing, biomedical nomenclature, out-of-vocabulary

1 Introduction

Familiarity with medical terminology assists medical practitioners and other stakeholders like doctors, nurses, and clinicians to understand rare and complex vocabulary. The evolution of medical terminology presents challenges in promoting the use of electronic health and medical records. For example, most medical terms originate from Greek and Latin words, making reading and spelling difficult Henderson and Dorsey (2019); Banay (1948). Medical researchers acquire conceptual skills with thorough learning of medical terms, dictionaries, and references such as Merriam-Webster Merriam-Webster (2018), WebMD WebMD (2012), and MedicineNet MedicineNet (2007), etc.

The electronic health records (EHRs) contain the diagnoses, pharmacological, and drug-disease concepts that provide a complete view of a patient's health. EHRs can inform drug discovery, treatment pathways, and real-world safety assessments. Unstructured text from EHRs can be encoded in a structured format (vectors) for downstream analysis using NLP methods. Unfortunately, the word embedding models faced the Out-of-vocabulary (OOV) words problem or used ineffective sub-word representations that caused low performance in intrinsic tasks to retrieve conceptual properties.

Popular embedding models including BERT (Devlin et al., 2019), ELMO (Peters et al., 2018), and FastText (Bojanowski et al., 2017) solve the OOV problem by using pre-processing tokenization techniques based on WordPiece (Wu et al., 2016), characters, and n-grams. These traditional NLP approaches are not built to understand the unique vocabulary and grammar of medical texts. For example, *mastodynia* is a disease whose meaning can be approximated from related and simple words like *breast, pain, and discomfort* rather than to approximate it with its non-logical sub-words or n-grams like [CLS],mast,##ody, and ##nia [SEP].

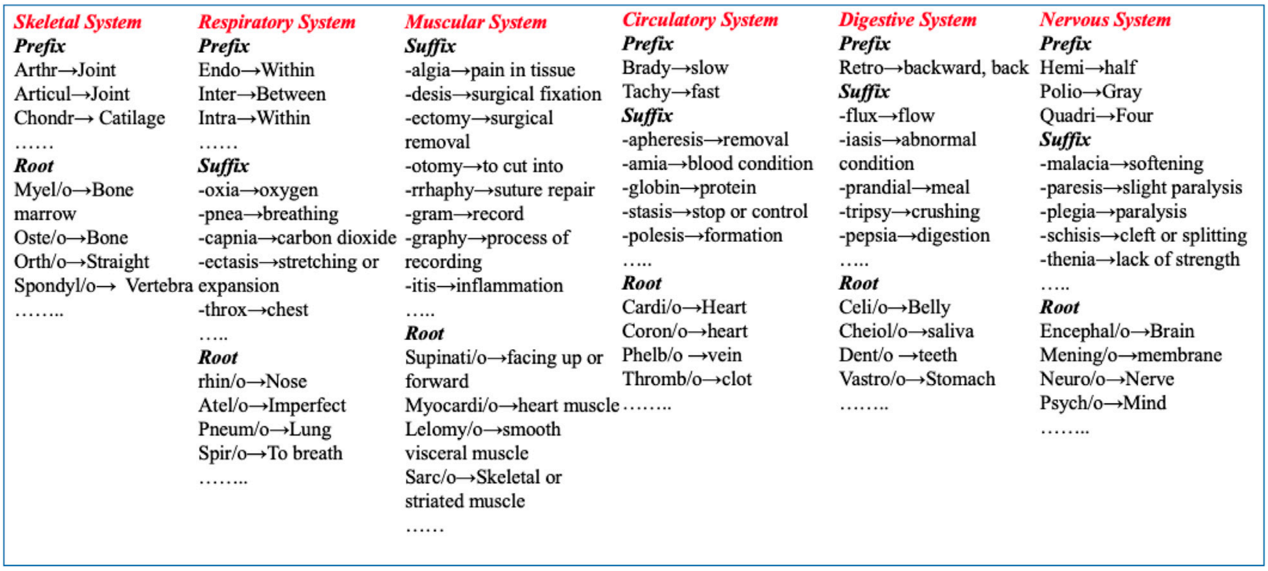
Biomedical and clinical terms have unique and complex characteristics such as prefixes, roots, suffixes, etc., therefore requiring a more focused effort around methodologies within the medical NLP domain (Banay, 1948; Meystre et al., 2008; Cohen and Demner-Fushman, 2014; Leaman et al., 2015; Henderson and Dorsey, 2019). In recent years, biomedical and clinical embedding models such as BioWordVec (Zhang et al., 2019) and BioNLP (Chiu et al., 2016) models have been trained under low capacity resource requirements like the Gensim library (Řehřek and Sojka, 2011). However, these models generally follow the Word2Vec (Mikolov et al., 2013a; Mikolov et al., 2013b) and GloVe (Pennington et al., 2014) algorithms, which face the OOV problem. The embedding models trained using the FastText algorithm (Bojanowski et al., 2017) claim to have solved the OOV problem, however they are ineffective.

The pre-trained embedding models generate either context-sensitive or distributed representations of word vectors. The context-sensitive models generate multiple embeddings for a word that capture the context based on its positional encoding learned using transformers or recurrent neural networks (RNN). Bidirectional Encoder Representations from Transformers (BERT) is a popular embedding model (Devlin et al., 2019), that has been extended to clinical and biomedical domains [ClinicalBERT Huang et al. (2019) and BioBERT Lee et al. (2020)]. These models tackle the OOV problem with the WordPiece algorithm (Wu et al., 2016) that represents a word by its frequent sub-words, e.g., immunoglobulin → (i,mm,uno,g,lo,bul,in). Embeddings from Language Models (ELMO) is another context-sensitive model that generates word-level embeddings using multiple convolutional neural networks (CNNs) with bi-directional LSTM (BiLSTM) (Peters et al., 2018). ELMO has also been extended to generate biomedical and clinical embeddings (Zhu et al., 2018; Jin et al., 2019; Subramanyam and Sangeetha. 2020). These studies deal with the OOV problem through character-level embeddings. Boukkouri et al. showed that character-level embedding was a better approach to removing biases in sub-words for biomedical terms than WordPiece e.g., choledocholithiasis → (cho,led,och,oli,thi,asi,s) (Boukkouri et al., 2020; Wu et al., 2016). The context-sensitive models are expensive, both in terms of computational and space resources since they train millions of hyperparameters with multiple attention heads.

The distributed representation models learn embeddings based on the word usage in a given corpus. The resultant vectors capture the contextual similarity between words. These static models generate a single vector per word and are trained either under Word2Vec (Mikolov et al., 2013b), GloVe (Pennington et al., 2014), or FastText (Bojanowski et al. (2017)). FastText enriches each word vector with its respective n-grams. It handles the OOV problem by leveraging the sum of n-gram vectors of the unknown word, e.g., $n = 3$, myocarditis $\rightarrow <$ my, myo, yoc, oca, car, ard, rdi, dit, iti, tis, is $>$. On the other hand, the embedding models trained by Word2Vec and GloVe face the OOV problem. These models replace unknown words with tags such as $<$UNK$>$ or a randomly generated vector, where different unknown words lose their uniqueness.

In this study, we proposed MedTCS, a novel medical terminology-based module that assists the pre-trained embedding models to generate vectors for unknown words and compound terms. It is an innovative post-processing solution that explores the given search space for those terms that are not directly present but whose semantic information is. MedTCS turns the word into its meaningful sub-words using the biomedical segmentation model. Ultimately, MedTCS helps the distributed representation models handle the OOV problem effectively.

We have compared MedTCS with recent state-of-the-art embedding models to investigate the effectiveness of capturing semantic information without encountering OOV problems. Our results showed that MedTCS enhanced the performance of pre-



Skeletal System
Prefix
Arthr→Joint
Articul→Joint
Chondr→ Catilage
……
Root
Myel/o→Bone marrow
Oste/o→Bone
Orth/o→Straight
Spondyl/o→ Vertebra
……..

Respiratory System
Prefix
Endo→Within
Inter→Between
Intra→Within
……
Suffix
-oxia→oxygen
-pnea→breathing
-capnia→carbon dioxide
-ectasis→stretching or expansion
-throx→chest
…..
Root
rhin/o→Nose
Atel/o→Imperfect
Pneum/o→Lung
Spir/o→To breath
……..

Muscular System
Suffix
-algia→pain in tissue
-desis→surgical fixation
-ectomy→surgical removal
-otomy→to cut into
-rrhaphy→suture repair
-gram→record
-graphy→process of recording
-itis→inflammation
…..
Root
Supinati/o→facing up or forward
Myocardi/o→heart muscle
Lelomy/o→smooth visceral muscle
Sarc/o→Skeletal or striated muscle
……

Circulatory System
Prefix
Brady→slow
Tachy→fast
Suffix
-apheresis→removal
-amia→blood condition
-globin→protein
-stasis→stop or control
-polesis→formation
…..
Root
Cardi/o→Heart
Coron/o→heart
Phelb/o →vein
Thromb/o→clot
……..

Digestive System
Prefix
Retro→backward, back
Suffix
-flux→flow
-iasis→abnormal condition
-prandial→meal
-tripsy→crushing
-pepsia→digestion
…..
Root
Celi/o→Belly
Cheiol/o→saliva
Dent/o →teeth
Vastro/o→Stomach
……..

Nervous System
Prefix
Hemi→half
Polio→Gray
Quadri→Four
Suffix
-malacia→softening
-paresis→slight paralysis
-plegia→paralysis
-schisis→cleft or splitting
-thenia→lack of strength
…..
Root
Encephal/o→Brain
Mening/o→membrane
Neuro/o→Nerve
Psych/o→Mind
……..

FIGURE 1
Understanding biomedical terms by mapping term components to human organ system.

trained models significantly in terms of coverage and/or semantic correlation. Moreover, we conducted experiments to assess the usefulness of enriched embedded vectors for downstream NER tasks (disease name identification and drug indication classification). MedTCS performed better than FastText in terms of performance and resource consumption on all tasks. The MedTCS module enhanced the performance of the FastText word vectors as compared to the n-gram and sub-word approaches used for unknown words (Flamholz et al., 2022). Furthermore, MedTCS is extensible with new terminologies and content.

2 Methodology

MedTCS is a lightweight module implemented in Python. It is a knowledge-driven system for forming terms by pluralizing, singularizing, and deconstructing words.

2.1 Meta-data collection

2.1.1 Word component dictionary

In MedTCS, we build meta-dictionaries for the prefixes, roots, and suffixes defining the meanings of medical term components. In addition to the lexical normalization and plural conversion of the unknown term, we have developed medical terminology-based look-up dictionaries for the parser by collecting information from "Medical Terminology for Dummies" (Henderson and Dorsey, 2019). The three semantic dictionaries contain 467 root words, 432 prefixes, and 112 suffixes, along with their corresponding meanings as shown in Figure 1.

2.1.2 Word segmenter model

MedTCS used Morfessor as a word segmenter model (Smit et al., 2014). In order to train the semi-supervised Morfessor 2.0 model, we used a corpora of 240 k words consisting of medical academic word list, e-biology, e-chemistry, and NLTK words (Bird and Loper, 2004; Wang et al., 2008).

2.2 MedTCS framework

Figure 2 provides a high-level description of our MedTCS module to encode OOV words from a set of sentences or words. In step (a), the OOV words are normalised for multiple morphological rules (represented as $N^{r}_{1,\ldots,n}$). In step (b), the remaining OOV words are exchanged with its plural or singular form by applying medical terminology-based rules (represented as $R^{r}_{1,\ldots,n}$). At each step, the normalized terms are encoded into vectors. In the succeeding steps (c) and (d), the words are passed to the parser, where dictionaries of prefix p, root r, and suffix s are used to tokenize them (represented as $P^{p,r,s}_{1,\ldots,n}$). Each component of the term is replaced with their respective meaning in the dictionary as a word list (represented as $M^{p_{w_i..w_k}, r_{w_i..w_k}, s_{w_i..w_k}}_{1,\ldots,n}$). The encoder encodes the tokens into its mean vector. Finally, the remaining non-encoded words are

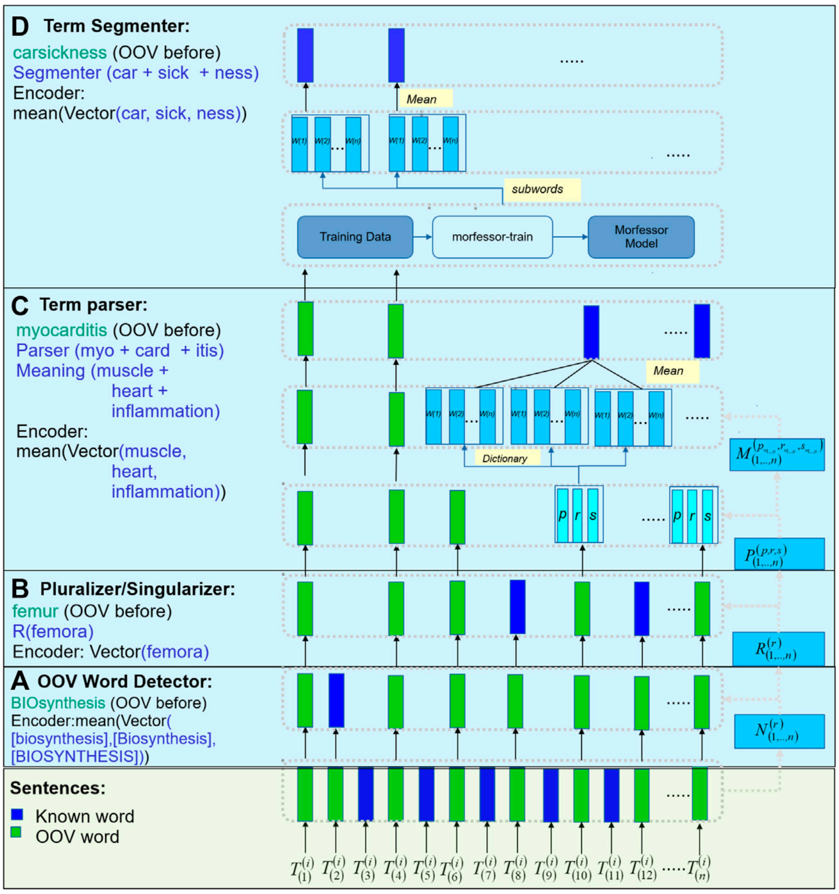

D Term Segmenter:
carsickness (OOV before)
Segmenter (car + sick + ness)
Encoder:
mean(Vector(car, sick, ness))
Mean
subwords
Training Data
morfessor-train
Morfessor Model
C Term parser:
myocarditis (OOV before)
Parser (myo + card + itis)
Meaning (muscle + heart + inflammation)
Encoder:
mean(Vector(muscle, heart, inflammation))
Mean
Dictionary
p r s
$M_{(1,\dots,n)}^{(p_{w_{1..n}}, r_{w_{1..n}}, s_{w_{1..n}})}$
$P_{(1,\dots,n)}^{(p,r,s)}$
B Pluralizer/Singularizer:
femur (OOV before)
R(femora)
Encoder: Vector(femora)
$R_{(1,\dots,n)}^{(r)}$
A OOV Word Detector:
BIOsynthesis (OOV before)
Encoder:mean(Vector([biosynthesis],[Biosynthesis], [BIOSYNTHESIS]))
$N_{(1,\dots,n)}^{(r)}$
Sentences:
🟦 Known word
🟩 OOV word
$T_{(1)}^{(i)}$ $T_{(2)}^{(i)}$ $T_{(3)}^{(i)}$ $T_{(4)}^{(i)}$ $T_{(5)}^{(i)}$ $T_{(6)}^{(i)}$ $T_{(7)}^{(i)}$ $T_{(8)}^{(i)}$ $T_{(9)}^{(i)}$ $T_{(10)}^{(i)}$ $T_{(11)}^{(i)}$ $T_{(12)}^{(i)}$ ..... $T_{(n)}^{(i)}$

FIGURE 2
MedTCS framework: **(A)** MedTCS detector normalizes the unknown terms and search in vocabulary; **(B)** Rule-based pluralizer or singularizer sub-module used to normalize the unknown terms; **(C)** Architecture for term-parser, where the compound words encode for its components that infer from the dictionary for its semantic words that encode as its mean vector; **(D)** Architecture for term segmenter, a pre-trained segmentation model segments the word into subwords that encodes as its mean vector.

TABLE 1 Statistics of Datasets.

Evaluation	Dataset	Corpus size	Type
Intrinsic Evaluation	UMNSRS-similarity Pakhomov et al. (2010)	566 term pairs	Pairwise similarity
	UMNSRS-relatedness Pakhomov et al. (2010)	588 term pairs	Pairwise relatedness
	MyoSRS Pakhomov et al. (2011)	101 term pairs	Pairwise relatedness
	EHR-RelB Schulz et al. (2020)	3630 term pairs	Pairwise relatedness
Extrinsic Evaluation	Dataset		
	BC5CDR Wang et al. (2019)	1500 articles	Disease Name
	NCBI-Disease Wang et al. (2019)	793 abstracts	Disease Name
	DICE Bhatt et al. (2021)	7231 sentences	Drug Indication

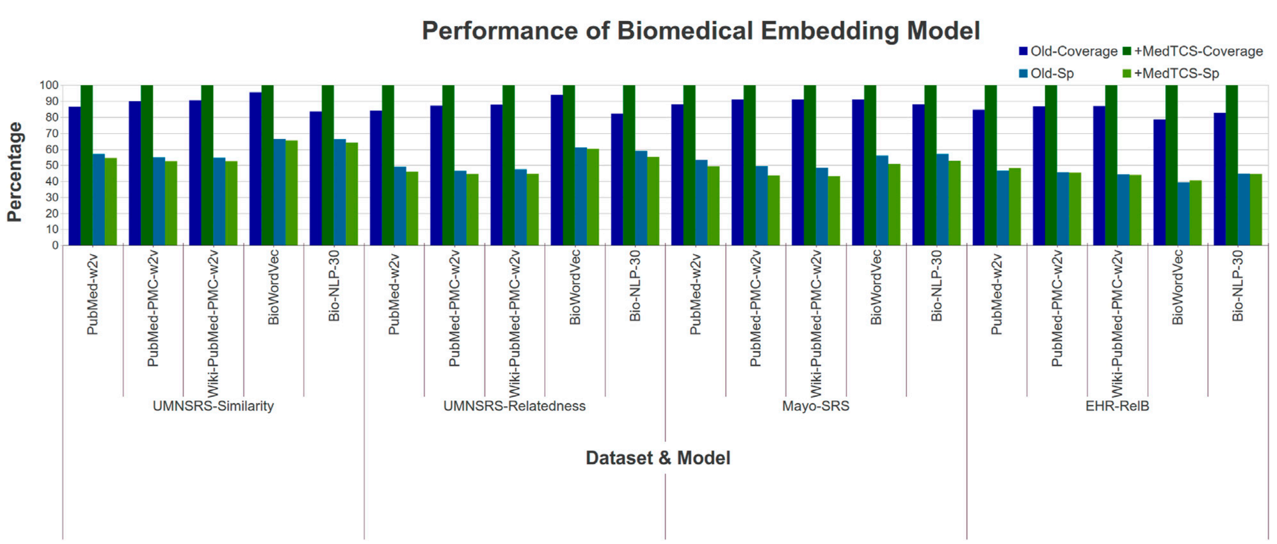


FIGURE 3
Comparison of performance variations in biomedical embedding model after adding MedTCS module on datasets of Table 1 for intrinsic evaluation.

passed to the pre-trained term segmenter model to intra-tokenize into meaningful words (that are also encoded as mean vectors).

2.2.1 MedTCS OOV word detector

The MedTCS OOV word detector identifies whether a token is known or unknown for a given vocabulary. The unrecognized word is passed through multiple normalization steps: 1) lexical property of the alphabetic case is applied, 2) intra-term punctuation marks are retained while ignoring starting and ending symbols, and 3) apostrophe symbols for OOV word detection are normalized.

2.2.2 MedTCS pluralizer/singularizer

The MedTCS pluralizer is based on the plural rules defined in medical terminology and implemented as a finite state machine. The sigularizer acts as a reverse finite state machine of the pluralizer.

2.2.3 MedTCS term parser

The MedTCS term parser was applied to an unknown word in two parts. First, the rule-based parser breaks the word into components of medical terminology, i.e., root, prefix, and suffix. Second, this parser implemented a dictionary lookup algorithm on each component to map its meaning. These dictionaries contained the definitions of the components of the medical terms collected from medical notes (Banay, 1948; Cohen and Demner-Fushman, 2014; Henderson and Dorsey, 2019). Each component in the dictionary belonged to one of the following human organ systems as shown in Figure 1 e.g., -pnea → breathing was a suffix belonging to the respiratory system. The root component is normalized for its combined form, like pneum/o → lung. Each component incrementally contributes in generating the vector representation of the unknown word. Each discovered vector by MedTCS term parser belonged to the lexical part of the unknown word and had attributes defined in the medical terminology. For example, *choledocholithiasis* → ["choledoch" (prefix)]+["o"]+["lithiasis" (suffix)] → [common bile duct]+[calculus or stone]. In case the term parser does not return a valid vector, the term segmenter was executed to determine meaningful sub-words of the unknown word.

2.2.4 MedTCS term segmenter

The MedTCS term segmenter is a wrapper around the Morfessor 2.0 module to acquire the meaningful sub-word units of an unknown term (Virpioja et al., 2013; Smit et al., 2014). We trained the system on a subset of Biology, Chemistry, and English corpora. Our word-level segmentation system returned the average vector of meaningful sub-words of an unknown term (like *seasickness* → sea + sick + ness).

2.3 Datasets

In addition to the widely tested UMNSRS similarity and relatedness datasets (Pakhomov et al., 2010), and the MyoSRS dataset (Pakhomov et al., 2011), our intrinsic evaluation included the latest and comparatively large benchmark named the EHR-RelB dataset Schulz et al. (2020). These datasets consist of word

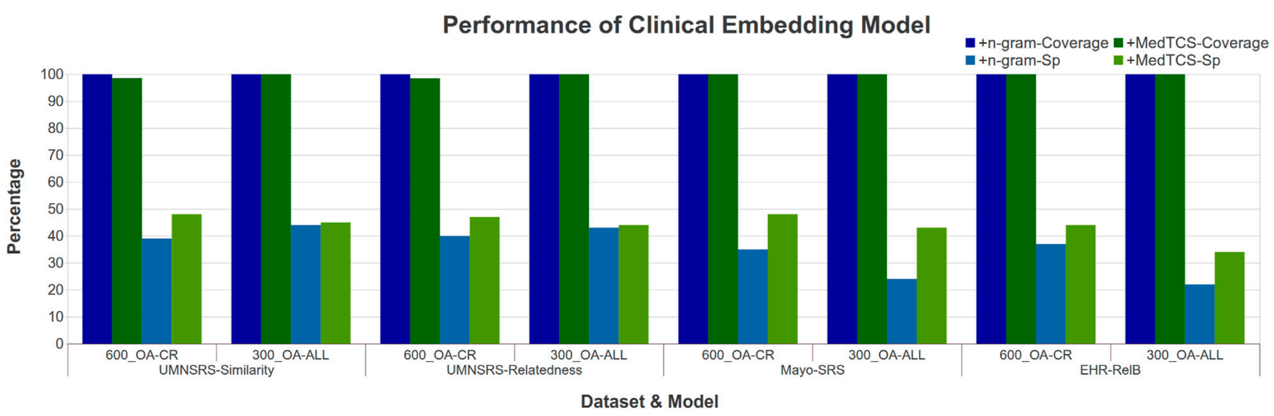


FIGURE 4
Comparison of performance variations in clinical embedding model after adding MedTCS module on datasets of Table 1 for intrinsic evaluation.

TABLE 2 Comparison of sub-word embeddings with word embedding + MedTCS on the UMNSRS-Similarity datasets.

Model		Version	Sp
BERT	BERT Devlin et al. (2019)	bert-base-uncased	0.07
	BioBert Lee et al. (2020)	dmis-lab/biobert-v1.1	0.30
	BlueBert Peng et al. (2019)	bionlp/bluebert_pubmed_mimic_uncased_L-12_H-768_A-12	0.36
	Bio_ClinicalBERT	emilyalsentzer/Bio_ClinicalBERT	0.23
Model	Alsentzer et al. (2019)	allenai/scibert_scivocab_uncased	0.18
	SciBERT Beltagy et al. (2019)		
	PubMedBERT Gu et al.(2022)	microsoft/BiomedNLP-PubMed BERT-base-uncased-abstract-fulltext	0.23
	CODER Yuan et al. (2022)	GanjinZero/UMLSBert_ENG	0.47
Word2Vec	PubMed-w2v	PubMed-w2v.bin	0.52
	+MedTCS		
	PubMed-PMC-w2v	//	0.49
Model	+MedTCS		
	Wiki-PubMed-PMC-w2v	//	0.49
+	+MedTCS		
	Bio-NLP-30 Chiu et al. (2016)	Bio-NLP-30	0.63
MedTCS	+MedTCS		
	BioWordVec Zhang et al. (2019)	BioWordVec	0.64
	+MedTCS		

pairs with their similarity or relatedness scores assigned by medical experts.

We checked the applicability of the MedTCS module to extract disease names from two publicly available datasets [NCBI-Disease and BC5CDR-Disease, Wang et al. (2019)] using the *BIO* scheme. *BIO* is used to encode entity annotations as token tags, where *B* indicates the beginning of the phrase, *I* is the element within the phrase, and *O* is the element outside of the phrase. Table 1 gives the details of benchmark datasets used for performance evaluation. We also used the Drug Indication Classification and Encyclopedia (DICE) dataset Bhatt et al. (2021) to check the performance enhancement achieved by MedTCS on classifying a sentence into indication or non-indication defined for five categories (indications, contradictions, side effects, usage instructions, and clinical observations). The dataset contained 7,231 sentences that were categorized into 4,297 indications, 1,673 clinical observations,

TABLE 3 Comparison of the word embedding + MedTCS best scores with latest reported results.

Model	UMNSRS-Similarity		UMNSRS-Relatedness		Model description
	#566	Sp	#587	Sp	
BioWordVec+ Graph Embeddings (GCN) Mao and Fung (2020)	480	0.629	473	0.590	A combined model of Graph convolutional network (GCN) a path-based graph embedding with BioWordVec embedding
Context2Vec+ BioWordVec+ PubMed + PMC Singh and Jin (2020)	471	0.634	484	0.561	Composite model of contextual embedding with BioWordVec concatenated with PubMed and PMC word embedding to achieve these results
CoderBERT Kalyan and Sangeetha (2021)	543	0.543	564	0.473	A BERT-based model obtained by fine-tuned a pre-trained BioBERT on UMLS synonyms and relations
SapBERT-S Kalyan and Sangeetha (2021)	543	0.585	564	0.505	A BERT-based model fine-tuned a pre-trained PubMedBERT on UMLS using a self-alignment objective to cluster the term concept
BioWordVec +MedTCS	**566**	**0.641**	**587**	**0.603**	BioWordVec with our composed MedTCS module, to extract the vector representation of a known and unknown term

Results with highest values of correlation and coverage scores are shown in bold.

701 contraindications, 492 usage instructions, and 68 side effects.

All the datasets discussed in Table 1 are publicly available in split form.

2.4 Evaluation metrics

In NLP, intrinsic evaluation extracts the semantic properties of pre-determined ground truth concepts with encoded vectors. On the other hand, extrinsic evaluation decodes the encoded information of embedding models and evaluates their efficiency in performing downstream tasks like NER. For the extrinsic evaluation, the coverage percentage is based on the number of encoded tokens of a dataset with the respective embedding model.

$$cosine\_similarity\,(A, B) = \frac{A.B}{\|A\|\times\|B\|}, \tag{1}$$

$$recall = \frac{TruePositives}{TruePositives + FalseNegatives}, \tag{2}$$

$$precision = \frac{TruePositives}{TruePositives + FalsePositives}, \tag{3}$$

$$F1\_score = \frac{2\times(precision\times recall)}{precision + recall}. \tag{4}$$

In intrinsic evaluation, the similarity scores are computed between the encoded term pairs using the cosine similarity as given in Eq. 1. Furthermore, these similarity scores are used with

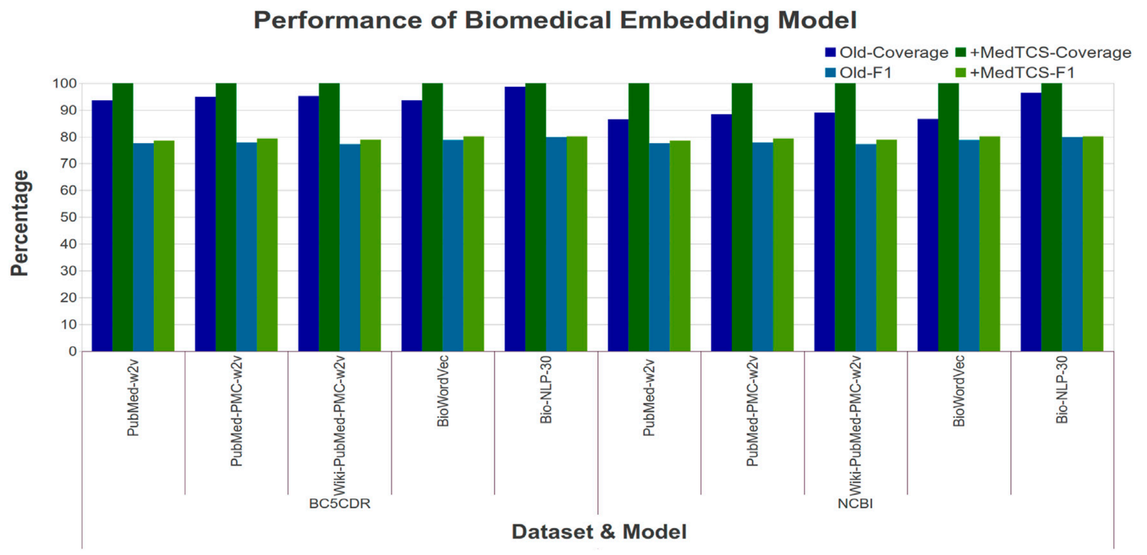


FIGURE 5
Comparison of performance variations in biomedical word embedding model after adding MedTCS module on datasets of Table 1 for NER task.

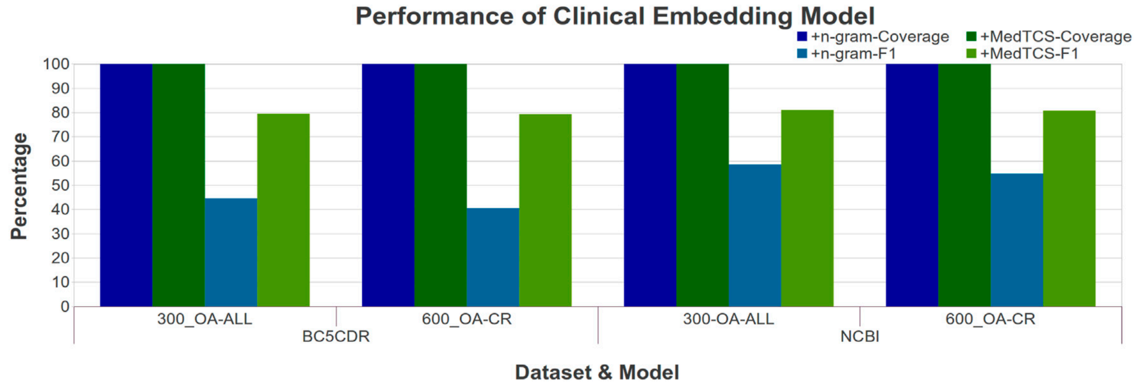


FIGURE 6
Comparison of performance variations in clinical FastText embedding model after adding MedTCS module on datasets of Table 1 for NER task.

the rankings by human experts to compute the Spearman (Sp) correlation coefficients with SciPy (Virtanen et al., 2020).

In extrinsic evaluation, the task of tagging the biomedical entities is performed by using a machine learning model trained on the encoded vectors. The performance measures used for this are recall, precision, and F1_score (Eqs 2–4).

3 Results

We compared the semantic and conceptual functionality of MedTCS with the pre-trained sub-word models (derivative models of BERT Devlin et al. (2019) and FastText Bojanowski et al. (2017)) for the biomedical and clinical domains.

3.1 Intrinsic evaluation

We evaluated the capability of MedTCS to enable the pre-trained word embedding models for encoding the OOV terms. MedTCS assisted the pre-trained word embedding models to achieve full coverage of all the conceptual pairs in the datasets. Moreover, we compared our model with related embedding models trained with FastText and BERT algorithms.

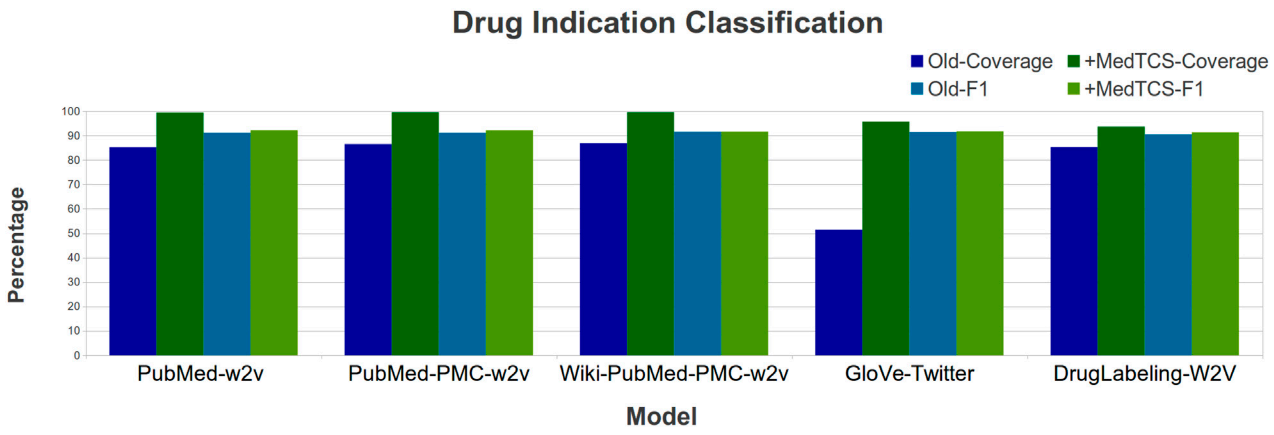


FIGURE 7
Model performances enhanced with MedTCS for Drug indication classification.

In this experiment, we included popular embedding models as a baseline, such as BioWordVec (Zhang et al., 2019), BioNLP (Chiu et al., 2016), PubMed-w2v, PubMed-PMC-w2v, and Wiki-PubMed-PMC-w2v (Moen and Ananiadou, 2013), most of which are defined under the Word2Vec algorithm (Mikolov et al., 2013b). Our baseline has the same encoder and decoder method for the NLP task without the MedTCS module. As the selected datasets include multi-word terms; therefore, the average vectors of each word are calculated with and without the MedTCS module.

Our analysis showed that the MedTCS module enabled all pre-trained embedding models to achieve full coverage with persuasive correlation scores on all datasets (Figure 3). For example, on the EHR-RelB dataset, the coverage of the *BiowordVec model* was enhanced from 2,857 terms pair to 3,630 terms pair and the Spearman (Sp) correlation also improved from 0.393 to 0.405. Overall, our results show that all models achieved 100% coverage of all the datasets with a slight decrease in correlation scores. As the OOV words are being approximated, therefore, a slight decrease in correlation scores is naturally expected.

We also enhanced the latest clinical word embedding models with MedTCS. The PMC Open Access Subset-Case reports (OA-CR) embedding models trained using word2vec/GloVe encountered the OOV word problem while working on the UMNSRS-Similarity dataset (Flamholz et al., 2022). MedTCS improved the coverage of all word embedding models from approximately 62% → 98% Supplementary Case S1.

Similarly, we analyzed the functionality of FastText to handle the OOV problem on the PMC Open Access subsets - Clinical Report (OA-CR) models and the PMC Open Access subsets - all manuscripts (OA-All) models (Flamholz et al., 2022). FastText trains each word vector along with its n-gram vectors. In the case of any OOV word, the average of its n-gram vectors are used to encode it (Bojanowski et al., 2017). For the FastText based OA-CR-600 embedding model, the Spearman (Sp) correlation value improved from 0.38 → 0.47 Supplementary Case S2. In conclusion, the MedTCS module enabled the different variants of OA-CR models to encode the vector for OOV terms from its search space effectively.

The OA-CR models have a small vocabulary; MedTCS enabled these models to achieve 100% coverage on all datasets as shown in Figure 4. Moreover, the MedTCS assisted the OA-CR models and the OA-ALL models to have significantly improved correlation scores, e. g., the FastText OA-All-300d model on the EHR-RelB dataset achieved 100% coverage and improved the Spearman (Sp) correlation scores from 0.25 → 0.35. The results on other variants of the OA-CR and OA-ALL embedding models for intrinsic evaluation are similar, as shown in Supplementary Figure S3.

On the other hand, BERT models use sub-words to solve the OOV word problem. We compared the proposed model with BERT and its derivative models defined for the clinical and biomedical domain [available on HuggingFace Wolf et al. (2019); Wolf et al. (2020)]. MedTCS outperformed BERT-based models by a significant margin in terms of correlation scores on the UMNSRS-Similarity dataset (Table 2). Moreover, in (Table 3), we have compared our best achieved results with recently reported scores of UMNSRS datasets (Mao and Fung, 2020; Singh and Jin, 2020; Yuan et al., 2022). MedTCS achieved significantly better coverage and correlation scores.

3.2 Extrinsic evaluation

Extrinsic evaluation requires training a system for the related downstream NLP tasks like NER, classification, etc,. The existing word embedding models achieve sub-optimal results due to the ineffective handling of OOV words (encoded unknown words with their n-gram vectors or a randomly generated vector). We

TABLE 4 Examples of the sub-word tokenization schemes followed by the different algorithms with the medical terminology-based MedTCS module.

Term	MedTCS	FastText	BioBert	CODER
		Bojanowski et al. (2017)	Lee et al. (2020)	Yuan et al. (2022)
mastodynia	breast, pain discomfort	<ma,mas,ast sto,tod,ody dyn.yni,nia,ia>	[CLS],mast,## ody,##nia, [SEP]	[CLS],mast,## odynia, [SEP]
prostatism	prostate, gland state,of,or,condition	<pr,pro,ros,ost,sta tat,ati,tis,ism,sm>	[CLS],pro,##sta ##tism, [SEP]	[CLS],prost,## atism, [SEP]
prostatorrhea	prostate, gland flow, excessive discharge	<pr,pro,ros,ost.sta tat,ato,tor,orr,rrh rhe,hea,ea>	[CLS],pro,##sta ##tor,##r,##hea [SEP]	[CLS],prost,## ator,##rh,##ea [SEP]
blepharospasm	eyelid,or,eyelash sudden,or involuntary	<bl,ble,lep,eph,pha har,aro,ros,osp,spa pas,asm,asm>	[CLS],b,##le,## pha,##ros,## pas,##m, [SEP]	[CLS],ble,## pha,#rosp,## asm, [SEP]
dyslipidemia	painful,fat,a blood, condition	<dy,dys,ysl,sli,lip pii,pid,ide,dem emi,mia,ia>	[CLS],d,##ys,## lip,##ide,## mia, [SEP]	[CLS] dyslipidemia [SEP]
dyspnea	painful, breathing	<dy,dys,ysp,spn pne,nea,ea>	[CLS],d,##ys,## p,##nea, [SEP]	[CLS],dyspnea [SEP]
urethrorrhea	urethra, flow excessive discharge	<ur,ure,ret,eth,thr hro,ror,orr,rrh,rhe hea,ea>	[CLS],u,##ret,## hr,##or,##r,## hea, [SEP]	[CLS],ureth,## ro,##r,##rh,## ea, [SEP]
arteriosclerosis	artery, hardening	<ar,art,rte,ter,eri rio,ios,osc,scl,cle ler,ero,ros,osi,sis,is>	[CLS],art,##eri ##os,##cle,## rosis, [SEP]	[CLS],arterio ##sc,##ler,## osis, [SEP]
dermatitis	Skin inflammation	<de,der,erm,rma mat,ati,tit,its,ts>	[CLS],der,##mat ##itis, [SEP]	[CLS] dermatitis [SEP]

tested the enriched vectors (by MedTCS) in identifying disease names from documents. We trained a bidirectional long-short term memory with a convolutional neural network (BiLSTM-CNN) Chiu and Nichols (2016) on the annotated corpus of BC5CDR and NCBI-disease (Table 1).

Figure 5 showed the performance enhancement in terms of coverage and F1 score (in percentage) achieved after replacing the randomly generated vector approach with our MedTCS module for OOV words. The MedTCS module with PubMed-w2v embedding enabled improved the coverage up to 13% on the NCBI-Disease dataset. Overall, MedTCS enabled word embedding models to achieve 100% coverage with an improved F1 score.

Similarly, we compared the MedTCS module with the n-gram approach for the NER task. Figure 6 showed that MedTCS improved the F1-score between 10 and 20% for the various embedding models as compared to the FastText n-gram vectors under the same parameters as for the BiLSTM NER system (Chiu and Nichols, 2016). The FastText OA-All-300d model with MedTCS achieved an F1-score of 0.80 (an improvement of 0.35) on the BC5CDR corpus and an F1-score of 0.81 (an improvement of 0.32) for the NCBI-disease corpus. Similar results were achieved on the other variants of the OA-CR and OA-ALL embedding models for the NER task (Supplementary Figure S4).

Bhatt et al. (2021) recently developed a Drug Indication Classification and Encyclopedia (DICE) based on FDA approved human prescription drug labeling. They also generated "DrugLabelling-W2V" embeddings based on Word2Vec and used them to classify each sentence into one of the five classes (indications, contradictions, side effects, usage instructions, and clinical observations). We enhanced the "DrugLabelling-W2V" embedding with the MedTCS module and improved the coverage by 9% and the F1_score by 1% (Figure 7).

4 Discussion

Curating a large corpus is the traditional approach in NLP to cover more concepts and enhance the vocabulary of word-level embedding models. For example, meta-data from dictionaries, meta-thesaurus, and hierarchical relationships from ontologies were also used as corpus. In the biomedical and clinical domains, the larger corpus of PubMed-PMC from MEDLINE and Wikipedia (Denoyer and Gallinari, 2006) was used to enlarge the vocabulary. Similarly, the NCBI sources, including the Medical Subject Heading (MeSH) (Lipscomb, 2000), the Unified Medical Language System (UMLS) metathesaurus concepts (NLM, 2004), and the Systemized Nomenclature of Medicine—Clinical Terms (SNOMED CT) concepts (Donnelly, 2006) have also been used as meta-corpus. The semantic content of the ontologies and the meta-data like Web Ontology Language (OWL) has also been used to train embedding vectors (Grau et al., 2008). In spite of these efforts, while encoding some rare terms and concepts, the embedding models still faced the OOV problem like in the BioWordVec embedding model (Zhang et al., 2019).

We have developed MedTCS, a module that generates the vector representation for unknown words based on medical knowledge. Different approximation techniques derived from medical knowledge bases have been used to encode the OOV words. To the best of our knowledge, this is the first-ever post-processing and run-time solution for the OOV problem that is specifically designed for pre-trained biomedical/clinical word embedding models. Each OOV word is parsed into its components, which are replaced with their meanings to generate the semantic vectors. In addition, MedTCS's segmentation model tokenizes compound words into its word units, as shown in Figure 2. The MedTCS module outperforms the FastText n-gram approach to handle OOV words as shown in Figure 4.

In an empirical analysis of the BERT and its derivative models, we have observed that these models can have a high cosine similarity value between pairs given in datasets (Table 1). However, in the task to measure the degree of contextual relatedness and similarity between biomedical and clinical terms, they showed decreased performance (Table 2). Furthermore, according to our findings on BERT models, CODER Yuan et al. (2022) has better performance, probably because it encodes most of the words without splitting them into their sub-words, as shown in Table 4.

Word embedding models are of great importance for various biomedical NLP applications, however they currently face a major problem of assigning vectors for unknown and rare words. To fill this gap, we have developed the MedTCS module to facilitate the pre-trained word representation models in encoding medical terms. We hope that our module will be considered as a standard medical term tokenizer for the application of NLP in the biomedical domain. MedTCS can also allow other biomedical NLP researchers to develop knowledge-based modules in a variety of real-world applications. Moreover, our research highlighted that there is a need to not only train large embedding models but also some knowledge-driven modules for the medical and clinical domains. According to our knowledge, MedTCS is the first post-processing and run-time solution for the OOV problem that improves the applicability and semantic efficiency of pre-trained embedding of medical terms.

Data availability statement

The code for our module and instructions for the use can be found on GitHub, https://github.com/NadiaSaeed/MedTCS.git. Publicly available datasets used for performance evaluation and analysis during the current study are available at https://github.com/cambridgeltl/MTL-Bioinformatics-2016. The original contributions presented in the study are publicly available. The generated dictionaries of MedTCS are available on GitHub, https://github.com/NadiaSaeed/MedTCS.git.

Author contributions

NS proposed and designed the study. NS and HN performed data analysis. NS and HN wrote and revised the manuscript. All authors read and approved the final manuscript.

Funding

This research work was funded by the Higher Education Commission of Pakistan and the Ministry of Planning

Development and Reforms under the umbrella of the National Center in Big Data and Cloud Computing (NCBC).

Acknowledgments

The authors acknowledge the support of Amna Binte Kamran, Hassan Mustafa, and Umair Ayub in proof reading the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.928530/full#supplementary-material

References

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., et al. (2019). "Publicly available clinical BERT embeddings," in Proceedings of the 2nd clinical natural language processing workshop (Minneapolis, MN: Association for Computational Linguistics), 72–78.

Banay, G. L. (1948). An introduction to medical terminology, Greek and Latin derivations. *Bull. Med. Libr. Assoc.* 36, 1–27.

Beltagy, I., Lo, K., and Cohan, A. (2019). "Scibert: a pretrained Language Model for scientific text," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing EMNLP, Hong Kong, China, November 3–7, 2019 (Association for Computational Linguistics)

Bhatt, A., Roberts, R., Chen, X., Li, T., Connor, S., Hatim, Q., et al. (2021). Dice: a drug indication classification and encyclopedia for ai-based indication extraction. *Front. Artif. Intell.* 4, 711467. doi:10.3389/frai.2021.711467

Bird, S., and Loper, E. (2004). *Nltk: the natural language toolkit*. Barcelona, Spain: Association for Computational Linguistics.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* 5, 135–146. doi:10.1162/tacl_a_00051

Boukkouri, H. E., Ferret, O., Lavergne, T., Noji, H., Zweigenbaum, P., and Tsujii, J. (2020). "CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters," in Proceedings of the 28th international conference on computational linguistics (Barcelona, Spain: International Committee on Computational Linguistics.), 6903–6915.

Chiu, J. P., and Nichols, E. (2016). Named entity recognition with bidirectional lstm-cnns. *Trans. Assoc. Comput. Linguist.* 4, 357–370. doi:10.1162/tacl_a_00104

Chiu, B., Crichton, G., Korhonen, A., and Pyysalo, S. (2016). "How to train good word embeddings for biomedical nlp," in Proceedings of the 15th Workshop on Biomedical Natural Language Processing, Berlin, Germany, August 7–12, 2016, 166–174.

Cohen, K. B., and Demner-Fushman, D. (2014). *Biomedical natural language processing*, 11. Amsterdam: John Benjamins Publishing Company.

Denoyer, L., and Gallinari, P. (2006). "The wikipedia xml corpus," in International Workshop of the Initiative for the Evaluation of XML Retrieval, New York, NY, June 1, 2006 (Springer), 12–19.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "BERT: Pre-training of deep bidirectional transformers forlanguage understanding," in Proceedings of NAACL-HLT 2019, June 2–7, 2019 (Minneapolis, MN: Association for Computational Linguistics), 4171–4186.

Donnelly, K. (2006). Snomed-ct: the advanced terminology and coding system for ehealth. *Stud. Health Technol. Inf.* 121, 279–290.

Flamholz, Z. N., Crane-Droesch, A., Ungar, L. H., and Weissman, G. E. (2022). Word embeddings trained on published case reports are lightweight, effective for clinical tasks, and free of protected health information. *J. Biomed. Inf.* 125, 103971. doi:10.1016/j.jbi.2021.103971

Grau, B. C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., Sattler, U., et al. (2008). Owl 2: the next step for owl. *J. Web Semant.* 6, 309–322. doi:10.1016/j.websem.2008.05.001

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., et al. (2022). Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Health* 3 (1), 1–23. doi:10.1145/3458754

Henderson, B., and Dorsey, J. L. (2019). *Medical terminology for dummies*. Hoboken: John Wiley & Sons.

Huang, K., Altosaar, J., and Ranganath, R. (2019). Clinicalbert: modeling clinical notes and predicting hospital readmission. *arXiv* [Preprint]. Available at: https://arxiv.org/abs/1904.05342.

Jin, Q., Dhingra, B., Cohen, W. W., and Lu, X. (2019). "Probing biomedical embeddings from language models," in Proceedings of the 3rd workshop on evaluating vector space representations for NLP (Minneapolis, MN: Association for Computational Linguistics), 82–89.

Kalyan, K. S., and Sangeetha, S. (2021). A hybrid approach to measure semantic relatedness in biomedical concepts. *arXiv* [Preprint]. Available at: https://arxiv.org/abs/2101.10196.

Leaman, R., Khare, R., and Lu, Z. (2015). Challenges in clinical natural language processing for automated disorder normalization. *J. Biomed. Inf.* 57, 28–37. doi:10.1016/j.jbi.2015.07.010

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., et al. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 1234–1240. doi:10.1093/bioinformatics/btz682

Lipscomb, C. E. (2000). Medical subject headings (mesh). *Bull. Med. Libr. Assoc.* 88, 265–266.

Mao, Y., and Fung, K. W. (2020). Use of word and graph embedding to measure semantic relatedness between unified medical language system concepts. *J. Am. Med. Inf. Assoc.* 27, 1538–1546. doi:10.1093/jamia/ocaa136

MedicineNet (2007). [Dataset]. Medical termsmedical dictionary, medical definitions medical terminology. Available at: https://www.medicinenet.com (Accessed 01 13.2022).

Merriam-Webster (2018). [Dataset]. Dictionary by merriam-webster America's most-trusted online dictionary. Available at: https://www.merriam-webster.com (Accessed 01 13.2022).

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., and Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb. Med. Inf.* 17, 128–144. doi:10.1055/s-0038-1638592

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv* [Preprint]. Available at: https://arxiv.org/abs/1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Adv. neural Inf. Process. Syst.* 26, 1–9. doi:10.48550/arXiv.1310.4546

Moen, S., and Ananiadou, T. S. S. (2013). "Distributional semantics resources for biomedical text processing," in Proceedings of LBM, Tokyo, Japan, December 12–13, 2013, 39–44.

NLM (2004). *Unified medical language system, rxnorm*. New Jersey, United States: National Library of Medicine. Available at:.

Pakhomov, S., McInnes, B., Adam, T., Liu, Y., Pedersen, T., and Melton, G. B. (2010). Semantic similarity and relatedness between clinical terms: an experimental study. *AMIA Annu. Symp. Proc.* 2010, 572. American Medical Informatics Association. doi:10.13020/D6CX04

Pakhomov, S. V., Pedersen, T., McInnes, B., Melton, G. B., Ruggieri, A., Chute, C. G., et al. (2011). Towards a framework for developing semantic relatedness reference standards. *J. Biomed. Inf.* 44, 251–265. doi:10.1016/j.jbi.2010.10.004

Peng, Y., Yan, S., and Lu, Z. (2019). . "Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets," in Proceedings of the 18th BioNLP workshop and shared task (Florence, Italy: Association for Computational Linguistics), 58–65.

Pennington, J., Socher, R., and Manning, C. D. (2014). "Glove: global vectors for word representation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, October 25–29, 2014 (EMNLP), 1532–1543.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., et al. (2018). . "Deep contextualized word representations," in Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: Human language technologies (New Orleans, LA: Association for Computational Linguistics), 2227–2237.

Řehřek, R., and Sojka, P. (2011). *Gensim—statistical semantics in python*. Paris: EuroScipy. *Retrieved from genism. org*.

Schulz, C., Levy-Kramer, J., Van Assel, C., Kepes, M., and Hammerla, N. (2020). . "Biomedical concept relatedness–a large ehr-based benchmark," in Proceedings of the 28th international conference on computational linguistics, December 8–13, 2020 (Barcelona, Spain), 6565–6575.

Singh, A., and Jin, W. (2020). "On using composite word embeddings to improve biomedical term similarity," in 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), October 26–28, 2020 (IEEE), 281–287.

Smit, P., Virpioja, S., Grönroos, S.-A., and Kurimo, M. (2014). "Morfessor 2.0: toolkit for statistical morphological segmentation," in The 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Gothenburg, Sweden, April 26-30, 2014 (Aalto University).

Subramanyam, K. K., and Sangeetha, S. (2020). Deep contextualized medical concept normalization in social media text. *Procedia Comput. Sci.* 171, 1353–1362. doi:10.1016/j.procs.2020.04.145

Virpioja, S., Smit, P., Grönroos, S.-A., and Kurimo, M. (2013). *Morfessor 2.0: python implementation and extensions for morfessor baseline*. Gothenburg, Sweden: Association for Computational Linguistics.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods* 17, 261–272. doi:10.1038/s41592-019-0686-2

Wang, J., Liang, S.-l., and Ge, G.-c. (2008). Establishment of a medical academic word list. *Engl. Specif. Purp.* 27, 442–458. doi:10.1016/j.esp.2008.05.003

Wang, X., Zhang, Y., Ren, X., Zhang, Y., Zitnik, M., Shang, J., et al. (2019). Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics* 35, 1745–1752. doi:10.1093/bioinformatics/bty869

WebMD (2012). [Dataset]. Webmd better information. better health. Available at: https://www.webmd.com (Accessed 01 13.2022).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2019). Huggingfaceś transformers: state-of-the-art natural language processing. *arXiv* [Preprint]. Available at: https://arxiv.org/abs/1910.03771.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). "Transformers: state-of-the-art natural language processing," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, November 16–20, 2020. 38–45.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., et al. (2016). Google's neural machine translation system: bridging the gap between human and machine translation. *arXiv* [Preprint]. Available at: https://arxiv.org/abs/1609.08144.

Yuan, Z., Zhao, Z., Sun, H., Li, J., Wang, F., Yu, S., et al. (2022). Coder: knowledge-infused cross-lingual medical term embedding for term normalization. *J. Biomed. Inf.* 126, 103983. doi:10.1016/j.jbi.2021.103983

Zhang, Y., Chen, Q., Yang, Z., Lin, H., and Lu, Z. (2019). Biowordvec, improving biomedical word embeddings with subword information and mesh. *Sci. Data* 6, 52. doi:10.1038/s41597-019-0055-0

Zhu, H., Paschalidis, I. C., and Tahmasebi, A. (2018). Clinical concept extraction with contextual word embedding. *arXiv* [Preprint]. Available at: https://arxiv.org/abs/1810.10566.