Heliyon 9 (2023) e16818

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Quality of word and concept embeddings in targetted biomedical domains

Salvatore Giancani [a,b], Riccardo Albertoni [b], Chiara Eva Catalano [b,*]

[a] Institut de Neurosciences de la Timone, Unité Mixte de Recherche 7289 Centre National de la Recherce Scientifique and Aix-Marseille Université, Faculty of Medicine, 27, Boulevard Jean Moulin, 13385 Marseille Cedex 05, France

[b] Istituto di Matematica Applicata e Tecnologie Informatiche, Consiglio Nazionale delle Ricerche, Via De Marini 16, 16149 Genova, Italy

ARTICLE INFO

Dataset link: https://github.com/SaGiancani/medical-concepts-embeddings

Keywords:
Embedding
Quality
UMLS
Coverage
Chronic obstructive pulmonary disease

ABSTRACT

Embeddings are fundamental resources often reused for building intelligent systems in the biomedical context. As a result, evaluating the quality of previously trained embeddings and ensuring they cover the desired information is critical for the success of applications. This paper proposes a new evaluation methodology to test the coverage of embeddings against a targetted domain of interest. It defines measures to assess the terminology, similarity, and analogy coverage, which are core aspects of the embeddings. Then, it discusses the experimentation carried out on existing biomedical embeddings in the specific context of pulmonary diseases. The proposed methodology and measures are general and may be applied to any application domain.

1. Introduction

Since the Mikolov's seminal work [25], embeddings have showed huge potential in several applications of natural language processing (NLP) tasks and have become part of the NLP toolkits. Moreover, pre-trained embeddings can be reused to build applications and inject knowledge in applications and intelligent systems. In the biomedical domain, embeddings are widely used in clinical tasks: clinical abbreviation expansion, text classification, named-entity recognition, information retrieval, clinical predictions, relation classification, de-identification of electronic health records, patient similarity.

The paper focuses on static embeddings. This can be seen as a limitation, considering the advent of contextual embeddings, which have proved to be more powerful in NLP applications. As a matter of fact, static embeddings continue to have worth and advantages, as discussed by Noh and Kavuluru [28]. First of all, better static word embeddings can also aid in the initialisation of embeddings to facilitate the process of language-modelling-based training of contextualised models. Moreover, simpler models that use static embeddings can be built with 1-2 orders of magnitude fewer parameters and can run on smaller CPUs even in low resource settings. Secondly, static embeddings can be of inherent utility for linguists to improve knowledge representation tools, for example, studying lexical semantics of biomedical language by looking at word embeddings and how they may be indicative of lexical relations (e.g., hypernymy and meronymy). Finally, contextualised embeddings are usually only useful in languages with extensive digital corpus. The language modelling aim on which such embeddings rely can result in considerable overfitting compared to static techniques, for less known languages with smaller repositories.

* Corresponding author.
E-mail address: chiara.catalano@ge.imati.cnr.it (C.E. Catalano).

https://doi.org/10.1016/j.heliyon.2023.e16818
Received 14 April 2023; Received in revised form 29 May 2023; Accepted 30 May 2023

Available online 2 June 2023

Moreover, word and concept embeddings play a fundamental role in representing the domain of interest from corpora, extracting the main concepts and serving relatedness among them. The main idea behind is the *distributional semantic hypothesis* [16]: terms such as words and concepts that are used and occur in the same contexts tend to exhibit related meanings. According to this principle, an embedding is a model representing terms into a space: each term of an n-dimensional vocabulary learned by the model is represented by a vector of m dimension; semantically similar terms have close distance in the space. In this perspective, static embeddings may offer an interesting basis for representing conceptual spaces [18], mediating and providing a complementary geometric representation to the symbolic or connectionist approaches largely adopted in cognitive systems. Accordingly, their interpretation as an alternative knowledge representation tool is still valuable and is the one we considered in this paper.

Kalyan and Sangeetha [20] reviewed a big amount of works, presenting and comparing popular embeddings for experimental usage and even for benchmarking. However, they neither evaluate how the reviewed embeddings cover the targetted domain of interest nor consider the geometry of domain-specific terms in the space. Geometrical features of the space influence the ability to support qualitative reasoning such as similarity, relatedness or analogy, which are important tools for practitioners. An ideal embedding should preserve domain-specific qualitative reasoning. Indeed, evaluating the available embeddings and choosing the best representing a certain domain of knowledge and not only a specific downstream applicationis an open research question [1][12]. Quality metrics and methodological frameworks to inspect existing embeddings are pivotal for promoting awareness and ad-hoc testing when reusing third parties embeddings, and impact on the trustability and robustness of the system that reuses the embeddings [14].

This paper provides a methodology to evaluate task-independent quality of biomedical embeddings. *Quality* is usually a multi-faceted concept that consists of different aspects, traditionally known as quality dimensions. This paper aims to address in particular the “coverage” dimension of embedding. The proposed method focuses on intrinsic quality of embeddings rather than extrinsic (i.e., downstream task-dependent) one. Although they are not always absolutely and immediately correlated, the intrinsic quality is an equally important component of the quality evaluation: firstly, when choosing third-party embeddings, there is not always a precise targetting application; secondly, even if there were, the targetting application is not necessarily the same taken into account in the downstream evaluation offered by others. In these scenarios, using both intrinsic quality measures may be more reliable.

We started from the assumption that the more an embedding returns elements coming from a specific domain, the better it covers that domain. The assumption is not restrictive and, in the end, gives a clear picture of how the embedding is able to reflect a determined knowledge. The main idea is targetting not only the terminological coverage, that is *how many* concepts of the domain of interest the embedding includes, but also *how well* the embedding supports *similarities* and *relatedness* within the domain. Moreover, the analogy and the *analogical reasoning* [25] [24] are addressed as well, encompassing the complete spectrum of the analysis. To the best of our knowledge, we are the first to deal with this definition of coverage for a specific domain.

As well as quality, the identification of the domain of interest lies in the eye of the beholder. For this reason, the methodology does not provide a fixed and rigid definition of what constitutes a domain, since it depends on the specific evaluation goal. Rather, the methodology operationalises the exploration of the domain of interest, relying on existing conceptualisations. In this way, it enables evaluators to tailor the coverage evaluation to their current needs.

We demonstrate our approach in the frame of pulmonary diseases, in particular, we focus on the Chronic Obstructive Pulmonary Disease (COPD). We re-purpose some state-of-the-art measures deriving from the information retrieval field to detect the presence of specific knowledge in the form of concepts and relations; we instantiate our approach to evaluate the coverage of the COPD domain among several medical word and concept embeddings. The embeddings were selected based on their availability, and considering a variety of corpora: they span from unstructured to structured knowledge, from informal to technical sources, from textual embeddings to the ones defined on Concept Unique Identifiers (CUIs) (see Section 3). Overall, the experimentation proved that the state-of-the-art embeddings only partially cover the concepts, similarities and analogies expected in the context of chronic obstructive pulmonary disease. The paper contribution is, thus, multifaceted. First, it emphasises evaluation and quality when reusing third-party embeddings; second, it promotes an evaluation methodology; and third, it assesses the coverage of current embeddings in the context of pulmonary diseases by carrying out a deep experimentation.

The provided methodology is general and may be applied to any specific domain, even not biomedical. The main prerequisite is the existence of a Knowledge Organisation System (KOS) to encode the domain of knowledge and the underpinning ground truth. In this paper we rely on UMLS, since it is a well-known and consolidated framework. Depending on the domain of interest, others may be considered: more and more knowledge graphs have been emerging, for example, KG Linking Open Drug Data [33], COVID-19 Knowledge Graphs [40] and Bio2RDF [7]. Whenever no domain-specific KOS is available, even general-purpose Knowledge Graphs, e.g. DBpedia [22] and Wikidata [38] may serve the purpose.

The article is structured as follows. Section 2 is devoted to the analysis of the state of the art related to the quality of the embeddings, while Section 3 describes thoroughly the proposed methodology and measures. Section 4 examines the results obtained in the experimentation runs, and Section 5 concludes the paper and discusses future directions.

2. Related work

Different works have considered the quality of embeddings from a domain-neutral perspective (e.g., [35,39,15]). Others have focused specifically on embeddings in the context of the biomedical domain [20,10,42,9]. The literature distinguishes between intrinsic evaluations and extrinsic evaluations. This paper contributes to intrinsic quality evaluation, which usually, “looks at how well the induced embeddings are able to encode syntactic and semantic information” [20]. The present work complements intrinsic

evaluation with metrics to assess if concepts, similarities and analogies relevant to a specific domain of interest are represented in embeddings.

We introduced the coverage of embeddings for a specific domain of interest as an explicit quality dimension. Some previous works relate to coverage to a certain extent. For example, facetE [17] proposes a methodology to build a benchmark against which it is possible to measure how sensitive embeddings are to eight general domains named "facets" (i.e., Technology, Geographic, Music, Movies and Game, Literature, and Economy). However, the specificity of the considered domains is different: FacetE inspects the sensitiveness to a set of very broad and generic facets, not related to specific medical domains. Moreover, the two approaches rely on different ground truths: while facetE derives the ground truth from the analysis of web available tables related to the various topics, this paper relies on the ground truth derived from a specialised metathesaurus, i.e. UMLS [4].

Other works propose quality metrics based on different kinds of knowledge organisation systems, such as ontologies and knowledge graphs. Alshargi et al. [1] extends the state of the art by providing several intrinsic metrics for evaluating the quality of embeddings learned from RDF Knowledge Graphs. It evaluates embeddings for their capability to represent structural aspects of the ontology underpinning the knowledge graph: categorisation, hierarchy, and relations. These measures are built upon different assumptions: (i) they do not deal explicitly with the coverage dimension; (ii) contrary to many embeddings considered in this paper, the adopted measures assume that the embeddings are trained on a knowledge graph, explicitly structured with an ontology.

The following part revises the related work and discuss the contribution of this paper in relation to the typical examples of intrinsic evaluation: Word Similarity and Relatedness tasks, Nearest Neighbour Search (NNS) and Word Analogy.

Word similarity and relatedness evaluations [23,42] check the proximity between the cosine similarity in the embedding space and reference similarity-relatedness scores. The reference scores either are in form of datasets of word (or concept) pairs along with their similarity (or relatedness) assigned by medical experts (e.g., UMNSRS Similarity and UMNSRS Relatedness [29], MayoSRS [30]) or are derived by existing Knowledge Organisation Systems like metathesauri, ontologies, knowledge bases (e.g. UMLS [4]). This paper does not assess embeddings against any reference similarity/relatedness scores, since building human-annotated scores for specific domain of interest like COPD is costly [10], and knowledge organisation systems are often not specific enough to define reliable similarity scores in these contexts. Nevertheless, it proposes a methodology to characterise the domain of interest in terms of seed concepts and seed relations derived by UMLS. Also, the aim is not to verify the perfect similarity among terms, but rather check if embeddings relate the terms relevant for a specific domain of interest: the more the embedding-induced similarity search involves elements coming from a specific domain, the better the embedding covers that domain.

Nearest Neighbour Search (NNS) evaluates the similarity by looking at the K-Nearest Neighbours of each concept in a particular embedding space to see if they belong to the same concept group (also known as class of success) as referenced in ontologies or lexicons [10]. Exploiting the NNS, Choi et al. [11] introduced evaluation metrics for validating the conceptual similarity and relatedness and the Medical Conceptual Similarity measure using UMLS semantic types as the class of success. The similarity coverage presented in this paper is inspired by the metric introduced by Choi et al. [11]. In particular, the measure relies on the Discounted Cumulative Gain (DCG) to measure the number of domain-specific terms returned by the embeddings. Differently from [11], the measure considers the whole set of terms related to the domain of interest (i.e., COPD) as the class of success for the DCG.

Word analogy verifies whether analogy reasoning on specific relations holds in the embedding or not. Mikolov et al. showed the ability of Word2Vec to support analogies such as "man is to woman as king to queen". The underpinning idea is that the analogy holds in an embedding if the vectors "man - king" and "woman - queen" share similar direction for an analogy relation (e.g., "is a"). Distinct formalisations have been provided, which differ in the way the optimisation problem is posed (e.g., 3CosAdd, PairDirection, 3CosMul [24]). However, if knowing three terms of the analogy, it is possible to check the fourth term by exploiting the vector arithmetic and cosine similarity.

The analogy coverage proposed in this paper measures the number of analogies relevant for the domain of interest that are supported by an embedding. The relevant analogies are extracted by UMLS considering the analogy that involves domain-specific concepts. The measure is built on top of 3CosAdd [24], and returns the percentage of domain-specific analogies supported by the embeddings.

3. Methodology

In the proposed methodology different components can be distinguished. A *domain representation*, such as a thesaurus, a metathesaurus, or an ontology that serves as encoding of the domain of knowledge and the underpinning ground truth in the domain. The *domain of interest* is the portion of domain for which we ideally test the appropriateness of the embeddings (e.g., COPD). Then, we introduce *seeds*, which are the collections of representatives of the domain of interest. Seeds are extracted from the domain representation and expected to be included in the embeddings. They express the concepts, relatedness among the concepts, and analogies to be tested, which are typically a subset of those available in the domain representation. Finally, the *metrics* score the presence of the seeds in the embeddings. In this paper, the UMLS metathesaurus is favoured for the domain representation, and different types of seeds and measures are defined, each of which is designed to answer a specific question about coverage:

Vocabulary coverage: Are the concepts from the domain of interest included in the embedding?
Similarity coverage: Are the concepts of the domain of interest represented as close in the space of the embedding?
Analogy coverage: Does the embedding support the analogical reasoning in the domain of interest? In other words, how well are the UMLS relationships related to the specific domain supported inside the embedding?

We considered most of the embeddings mentioned in [20]: the ones publicly available, based on either text or CUIs. They are all classical static models, which are not explicitly trained on sentences or paragraphs, then the vocabularies are usually built either on short multi-words composed terms or on one-word terms. Some static word embeddings (e.g., FastText [5] and staticised contextual embeddings [6]) have a direct strategy for solving the problem of concepts expressed by multiple words, but not all architectures provide a solution for it. Some embeddings build upon CUIs provided by UMLS instead of words (e.g., [11,3,13]): we included them as CUI embeddings in our analysis.

Section 3.1 introduces the reasons that have brought to the selection of UMLS as domain representation. Section 3.2 describes how the seeds for the different types of metrics are extracted from the domain representation. Section 3.3 defines the metrics for each kind of coverage.

3.1. Domain representation and ground truth

We choose UMLS as a reference for the domain representation, ground truth and as a source for characterising the domain of interest. Indeed, the usage of a biomedical knowledge base assures robustness and preciseness, otherwise reachable via experts' human assessment. Other sources, such as open-source datasets constituted by pairs of similar concepts, are often used as ground truth to evaluate proportions among concepts [31][30] or may be suitable for other biomedical NLP tasks [37][43]. However, they poorly cover the COPD domain and are unusable as specific biomedical domain representations. For example, the reference standard dataset proposed by Pedersen et al. [31] shows only one pair of words related to "Chronic obstructive pulmonary disease" on a dataset of 30 pairs. In the dataset used by Pakhomov et al. [30], only one pair involves explicitly "Chronic obstructive pulmonary disease", two pairs "Pneumonia", and no one "Asthma", on a total of 101 pairs. Adopting UMLS as domain representation and source from which to build the domain of interest provides two main advantages: it provides a standard set of terms, synonyms and identifiers to use when querying the embeddings in the biomedical domain; it provides a domain specific ground truth, preventing from relying on an extended annotation performed by experts, which has to be built on purpose and on a very limited domain of interest, supplying the scarcity of ground truths on such a domain. Moreover, UMLS is a consolidated framework: a wide range of tools for the normalisation of concepts or even for CUIs to words -and vice versa- conversion are maintained [2][34].

3.1.1. Unified medical language system

According to the system documentation available online, UMLS "is a comprehensive collection of multilingual controlled vocabularies in the biomedical sciences, including the International Classification of Diseases and Related Health Problems (ICD-10), Medical Subject Headings (MeSH). It is built from lists of controlled terms used in patient care, health services billing, and biomedical literature and structured in concepts" [4]. Each concept has a unique and permanent concept identifier (CUI) and a set of terms or strings providing the lexical representations for the concept. Among the lexical representations, one of them is selected as the preferred label, the other terms provide lexical variation and synonyms used in documents or other sources in view of the preferred label. Apart from the lexical variations for concepts, UMLS provides non-synonymous relationships between concepts from the same source vocabulary and between concepts in different vocabularies. All relationships carry a general label (REL), describing their basic nature, such as Broader, Narrower, Child of, Qualifier of. Some of them also carry an additional label (RELA) that explains the nature of the relationship more precisely. We focused on RELA labels for our evaluation of relationships, as shown in the next sections. Moreover, all concepts are assigned at least one semantic type.

3.2. Seeds

Seeds represent the entities that an embedding should distinguish according to the UMLS characterisation. The methodology defines two different types of seeds: seeds for concepts, which consist of a set of representative concepts of the domain of interest, and seeds for analogies, which consist of concept pairs linked by relationships in the domain of interest.

3.2.1. Seeds for concepts

The seeds for concepts are built using two different strategies. In the first, UMLS is queried to obtain a list of concepts at a one-hop distance by the COPD concept, uniquely identified by a CUI code (e.g., 'C0024117'): with one-hop we mean a distance of one relationship away. The total number of seeds are in this case 256: in other words, this number represents all the possible unique concepts directly related to COPD inside UMLS. We call this set *seeds by UMLS relations*. A second strategy consists of extracting a set of concepts from a plain text, fed to the MetaMap [2] conversion tool, which maps concepts from text to UMLS CUIs. The chosen text is [8], given the wide variety of COPD-related aspects faced and the high number of citations. This strategy returns several hundreds of concepts, since qualitative concepts and typical terminology of natural language are included: the less significant concepts were discarded, according to the MetaMap ranking system, choosing the first 399 concepts. We call this second strategy *seeds by MetaMap*. In both cases, seeds are represented as a set of CUIs: the final *seeds for concepts* is the union of the two sets.

3.2.2. Seeds for analogies

Analogical reasoning requires concept pairs, connected by specific relationships: similarly to $\overrightarrow{Queen} - \overrightarrow{King} = \overrightarrow{Woman} - \overrightarrow{Man}$ [25] with the dualism role/genre and a simple "is a" relation, in a biomedical case we could have $\overrightarrow{Cycloserine} - \overrightarrow{Acetylcysteine} = \overrightarrow{Tuberculosis} - \overrightarrow{Bronchitis}$, with a dualism drug/disease and a medical relation "treats".

Table 1
List of the considered UMLS relationships.

associated_finding_of	has_associated_finding
associated_morphology_of	has_associated_morphology
associated_with_malfunction_of_gene_product	gene_product_malfunction_associated_with_disease
clinical_course_of	has_clinical_course
contraindicated_with_disease	has_contraindicated_drug
course_of	has_course
disease_has_associated_anatomic_site	is_associated_anatomic_site_of
disease_has_associated_gene	gene_associated_with_disease
finding_site_of	has_finding_site
manifestation_of	has_manifestation
may_treat	may_be_treated_by

For the analogical case, we defined the set as in eq. (1):

$$R_{UMLS} = \{R_i : (COPD\ R_i\ x) \vee (x\ R_i\ COPD), x \in UMLS\} \tag{1}$$

where R_i are a subset of the relationships in UMLS. Only medical relationships are kept: identity relations, relations about UMLS system versioning, the empty relation and more general relations (e.g., inverse_isa, isa) have been discarded. We considered most relationships used by [9]; a few were discarded because of the different case studies (e.g., has_ingredient). [9] includes 6 drug-related relationships, only in one direction; contrarily, we consider more relationships and also the correspondent inverse one for each of them. UMLS provides a set of 54 relationships in which the COPD concept is involved. After polishing, we selected 22: 11 relations and their inverse ones. The table (1) shows the considered relations.

We define the *seeds for analogies* for the relation $R_i \in R_{UMLS}$ as the set in eq. (2):

$$W_i = \{(x, y) : (x\ R_i\ y) \wedge ((x \in seed) \vee (y \in seed)) \mid R_i \in R_{UMLS}\} \tag{2}$$

where $seed$ is *seed by UMLS relations*. We opted for *seeds by UMLS relations* instead of *seeds by MetaMap* since the second may give more general concepts and we preferred to avoid pairs which are too distant from the COPD domain. Moreover, being bigger, the second is computationally expensive.

3.3. Measures

The metrics score the presence of the seeds in the embedding. We considered both word embeddings and CUI embeddings. Consequently, metrics should act on both. We used UMLS to convert the seeds according to the type of embedding.

The *seeds for concepts* is represented by a set of CUIs, and it is used for vocabulary coverage and similarity coverage. In case of CUI embeddings [11][13][3], no specific mapping between the seeds and the embedded vocabulary is needed. In the case of word embeddings, instead, the embedded vocabulary includes lexical representations, such as words, and needs to be mapped into the seeds to work out the metrics. We exploited the mapping between lexical representations and CUIs provided by UMLS. UMLS associates a list of textual labels to each CUI, either built ex-novo or emerging from a rich corpus of scientific literature. It identifies a preferred label for each concept, but the preferred label is rarely represented inside embeddings. As a remedy, the UMLS best-ranked lexical variations found inside the embedding vocabulary are preferred, maximising the presence of seeds inside a particular embedding.

This mapping strategy is a prerequisite for the application of the measures related to the In-Vocabulary IV sets (eq. (4) and eq. (10)) and Out-Of-Vocabulary OOV sets (eq. (3) and eq. (11)) and consequently $\%CG$ (eq. (6)) and $posDCG$ (eq. (8)).

The *seeds for analogies* (eq. (2)) are used in analogy coverage and are composed by a set of couples, where at least one of the two concepts in the couple belongs to the *seeds by UMLS relations*. The analogical reasoning requires two couples for closing the analogy: the quadruples are built as a partial permutation of couples from the *seeds for analogies*. The same conversion strategy adopted for the seeds for concepts has been applied to the seeds for analogies. In the case of *seeds by UMLS relations*, a concept can be inside or out of vocabulary of an embedding; for the *seeds for analogies* we consider as out of vocabulary those quadruples where at least one of the 4 elements is out of vocabulary. This allows us to define the In-Vocabulary IV_i^{ar} (eq. (10)) and Out-Of-Vocabulary OOV_i^{ar} (eq. (11)) sets for analogy coverage, which are prerequisites for the application of the analogical reasoning metric AR_i (eq. (12)) and its normalised version M_i^{ar} (eq. (16)).

3.3.1. Vocabulary and similarity coverage

Vocabulary coverage regards the presence of the domain inside the embedding or, better, how much the embedding covers the domain. As previously indicated, seeds represent the interest domain, so it is natural to measure the vocabulary coverage in terms of the cardinality of seed set present in the embeddings. We distinguish different metrics for measuring the vocabulary coverage: the Out-of-Vocabulary (OOV), the In-Vocabulary (IV), and the percentage In-Vocabulary (%IV). The $|\,OOV\,|$ counts the number of seeds not included in the embedding (eq. (3)):

$$|\,OOV\,| = |\,seed \setminus V_{emb}\,| \tag{3}$$

where V_{emb} is the embedded vocabulary set.

Complementarily, the $|\,IV\,|$ counts the number of seeds included in the embedding vocabularies (eq. (4)):

$$|\,IV\,| = |\,seed \setminus OOV\,| \tag{4}$$

Hence, *how much* the embedding covers the domain of interest could be defined as the percentage value with respect to the seed set cardinality (eq. (5)):

$$\%IV = \frac{|\,IV\,|}{|\,seed\,|} \tag{5}$$

Analogously, we can define %OOV as the percentage value of eq. (3).

Similarity coverage focuses on the neighbourhood of a queried concept, and regards *how well* the immediate similarity-adjacent concepts reflect the domain of interest. For each concept in the seed, the more the embedding reflects terms of the domain in the concept k-Nearest Neighbours, the more the embedding is considered interesting for the domain. The k-NN algorithm allows to detect the k-most-similar elements according to the geometric space of concepts learned by the analysed embedding: indeed, the most similar elements are selected according to the cosine similarity among the elements and the chosen seed.

If the embedding is properly trained and covers adequately the analysed domain, a big number of elements among the chosen ks would be inside the seed, which ideally corresponds to the fact the embeddings learned meaningful proportions among embedded concepts. If a word or CUI embedding was trained on domain-related documents (scientific papers in the pulmonary field or even COPD patients' clinical records) an optimal vocabulary geometrical disposition would be obtained, according to COPD/pulmonary domain: in other words, for each seed concept, the closest k elements would be inside the seed. Oppositely, an embedding could be well trained with a large vocabulary size, but, if the training dataset is not domain centred, the risk would be not having the seed elements in the k-NN. The $\%CG$ is formalised as in eq. (6):

$$\%CG_k = \frac{1}{k\,|\,seed\,|} \sum_{i=0}^{|seed|} \sum_{j=1}^{k} \mathbb{1}_{seed}^{ij} \tag{6}$$

where $\mathbb{1}_{seed}^{ij}$ represents the *relevance score*, and is defined as in eq. (7):

$$\mathbb{1}_{seed}^{ij} = \begin{cases} 1 & j \in seed \\ 0 & j \notin seed \end{cases} \tag{7}$$

where j is the jth element of the list of the k most-similar elements to i in the embedding.

Using the normalisation described above, the metric returns the percentage of seeds contained in the set of k most similar elements. Further information of *relatedness* is provided by equations derived by the class of Discounted Cumulative Gain (DCG) [19] [41]. The DCGs are widely employed as a ranking measure, in several applications: the most powerful and immediate example is probably provided by search engines. The DCGs provide information even about the position of the k-most-similar elements, not only on the occurrence in the neighbourhood. We can use the DCGs to rank when a desired element or undesired element is found in the k-most-similar neighbourhood, naming the DCGs as Positive or negative Discounted Cumulative Gain, respectively. We formalise the positive Discounted Cumulative Gain (posDCG) as in eq. (8):

$$posDCG_k = \frac{1}{h\,|\,seed\,|} \sum_{i=0}^{|seed|} \sum_{j=1}^{k} \frac{\mathbb{1}_{seed}^{ij}}{\log_2(j+1)} \tag{8}$$

where h is the normalisation factor as in eq. (9):

$$h = maxDCG = \sum_{j=1}^{k} \frac{1}{log_2(j+1)} \tag{9}$$

which corresponds to the max value obtainable by the DCG computation, otherwise known as ideal DCG: in this way, eq. (8) returns a value between 0 and 1.

3.3.2. Analogy coverage

Analogy coverage conveys information about *how well* the investigated embedding expresses specific relationships supporting the analogy reasoning. Our analogy coverage grounds on the previous formalised sets W_i (eq. (2)), which represent the couples bound by a chosen relation R_i in UMLS. We introduce two metrics in analogy with the In-Vocabulary and Out-of-Vocabulary defined for similarity coverage: the In-Vocabulary and the Out-Of-Vocabulary for analogical reasoning.

$|IV_i^{ar}|$ counts how many seed couples are actually included in the embedding, being the presence of the UMLS-derived couples a prerequisite for supporting the relation in the embedding (eq. (10)):

$$|IV_i^{ar}| = \left|V_{emb}^2 \cap W_i\right| \forall i \tag{10}$$

where V_{emb} is the embedded vocabulary set.

A version of eq. (3) for the analogy coverage can be defined as in eq. (11):

$$|OOV_i^{ar}| = \left|W_i \setminus IV_i^{ar}\right| = \left|W_i\right| - |IV_i^{ar}| \forall i \tag{11}$$

The actual check that an analogy between two couples holds is scored computing algebraic operations among concepts (i.e., vectors) inside the embedding. The adopted approach is inspired by the metric initially appeared in [25] and mathematically formalised in [24] as $3CosAdd$: doing simple operations like addition or difference between vectors in an embedding space, we are able to obtain the hypothetical expected concept which closes the analogical reasoning. In the case of the analogy $\overrightarrow{Cycloserine} - \overrightarrow{Acetylcysteine} = \overrightarrow{Tuberculosis} - \overrightarrow{Bronchitis}$, the expression $\overrightarrow{Cycloserine} - \overrightarrow{Acetylcysteine} + \overrightarrow{Bronchitis}$ is supposed to correspond to $\overrightarrow{Tuberculosis}$. We ground our metric on the expected outcome: if a certain relationship is properly represented inside the embedding, the distance between the two concepts of each semantic couple related by that relationship will have same value, ergo the outcome would be the expected one. The Analogy Reasoning Coverage formalises this as in eq. (12):

$$AR_i = \sum_{h=0}^{|IV_i^{ar}|} \sum_{\substack{j=0 \\ j \neq h}}^{|IV_i^{ar}|} \mathbb{1}_{ar}^{hj} \forall i \tag{12}$$

With $\mathbb{1}_{ar}^{hj}$ defined as in eq. (13):

$$\mathbb{1}_{ar}^{hj} = \begin{cases} 1 & x_j \in \text{k-NN}(x_h - y_h + y_j) \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

$\mathbb{1}_{ar}^{hj}$ returns 1 when the expected value x_j is among the k vectors determined by the k-NN algorithm, 0 otherwise. The cosine similarity is used as distance metric for k-NN computation: every time the element which is expected to close the analogical expression is among the first ks, we consider contemplated the analogy for the two couples, with the indices h and j. The result of the analogical measure (eq. (12)) is a count of the analogical expressions contemplated for a specific relationship in a certain embedding. A normalisation factor could be applied to eq. (12) introducing the m_i^{ar} defined in eq. (15), which provides a relative percentage.

First, we compute all the partial permutation of couples in IV_i^{ar} set equal to (eq. (14)):

$$\frac{|IV_i^{ar}|!}{(|IV_i^{ar}| - 2)!} = |IV_i^{ar}| \left(|IV_i^{ar}| - 1\right) \tag{14}$$

Since IV_i^{ar} depends on the embedding vocabulary $Vemb$, the chosen relationship R_i and on the $seed$, an adjusting multiplicative factor has to be applied and we define m_i^{ar} as in eq. (15):

$$m_i^{ar} = \frac{|IV_i^{ar}| \left(|IV_i^{ar}| - 1\right)}{|W_i| \left(|W_i| - 1\right)} \forall i \tag{15}$$

The usage of $|W_i|$ -the number of all couples of UMLS for the relation i- is needed to compare measures among embeddings, being not dependent on any of the considered embeddings. Finally, we define the Percentage Analogy Coverage as in eq. (16):

$$M_i^{ar} = \frac{m_i^{ar}}{|IV_i^{ar}| \left(|IV_i^{ar}| - 1\right)} AR_i = \frac{AR_i}{|W_i| \left(|W_i| - 1\right)} \forall i \tag{16}$$

4. Experimentation

The carried out experiments evaluate most of the embeddings cited in [20] according to the proposed measures. The experiments considered the embeddings that explicitly encode words or concepts -CUIs-, discarding those not available or encoding different entities such as syllables, sentences, patients. Table 2 lists such embeddings with the cardinality of their vocabulary and the corresponding reference. The upper part of the table includes *CUI embeddings* with augmented information from UMLS knowledge base. The bottom part shows *word embeddings*. The application they were used for is also mentioned. The metrics code and the experimentation results are available at the github repository.

The experiments have been run to assess similarity coverage for different values of k. In the next, the results for only $k = 10$ will be shown, since it is convenient in many applications such as query expansion and similarity search. The seed considered in the experimentation is the union of *seeds of UMLS relations* and *seeds by MetaMap*.

Table 3 shows the results for the analysed embeddings. The comparison between the vocabulary dimension (V_{emb} in Table 2) and the vocabulary coverage metrics in Table 3 suggests that the cardinality of the embedded vocabulary is not a very indicative metric for the coverage per se. For example, embeddings having quite different vocabulary cardinalities such as Pubmed, Pubmed&PCM, wiki&Pubmed&PCM result in equal percentage In-Vocabulary coverage $|\%IV|$ in Table 3. The above embeddings include the same percentage of seeds despite the differences in their vocabulary magnitude.

Considering the vocabulary coverage measured by $\%IV$, the word embeddings score higher than CUI embeddings. The PubMed and PMC models show best and comparable results (between 71.90% and 72.21%); whereas cui2vec has the best performance among the CUI embeddings (48.51%). Only Healthvec and tweetsvec show lower vocabulary coverage than CUI embeddings. The PMC model is trained with a smaller corpus than the others, which partially share the training corpus: PubMed corpus is common to

Table 2
The embeddings targetted in the experimentation. The first three rows refer to CUI embeddings, while the others are word embeddings.

Embedding	Source	$\#V_{emb}$	Application	URL
claims_cuis	[11]	14852	Cohort selection and patient summarization	git
DeVine	[13]	52102	Medical information retrieval	see [11]
cui2vec	[3]	109053	Medical information retrieval	git
PMC	[26]	2515686	Text classification, named entity recognition and query expansion	website
PubMed		2351706		
PubMed&PMC		4087446		
wiki&PubMed&PMC		5443656		
Healthvec	[36]	73644	Drug assumption behaviour analysis via Social-Media text processing	git
GoogleNews	[25]	3000000	Text classification	website
tweetsvec	[27]	26278	Pharmacovigilance from social media	website

Table 3
Measures obtained for vocabulary and similarity coverage considering as set of seed the union of *seeds of UMLS relations* and *seeds by Metamap*. In the upper rows the CUI embeddings, word embeddings below.

Embedding	$\lvert OOV \rvert$	$\%OOV$	$\lvert IV \rvert$	$\%IV$	$\%CG_{10}$	$posDCG_{10}$
claims_cuis	**584**	**91.68%**	53	8.32%	0.005	0.007
DeVine	356	55.89%	281	44.11%	0.035	0.042
cui2vec	328	51.49%	309	48.51%	**0.045**	0.049
PMC	179	28.10%	458	71.90%	0.042	0.051
PubMed	177	27.79%	**460**	**72.21%**	0.041	**0.053**
PubMed&PMC	177	27.79%	**460**	**72.21%**	0.043	0.052
wiki&PubMed&PMC	177	27.79%	**460**	**72.21%**	0.039	0.047
Healthvec	381	59.81%	256	40.19%	0.019	0.019
GoogleNews	203	31.87%	434	68.13%	0.023	0.029
tweetsvec	416	65.31%	221	34.69%	0.020	0.022

PubMed, PubMed&PMC, wiki&PubMed&PMC models, and this is likely the reason why these models share large part of the inside-the-vocabulary domain concepts. Although GoogleNews is trained with general information, it outperforms the CUI embeddings, while showing a performance significantly lower than PubMed and PMC resources. To give an indication, Table 4 shows examples of COPD-related concepts and whether or not they are found in the embeddings.

Contrarily to vocabulary metrics, similarity coverage does not show a clear cut in performances between CUI and word embeddings: the maximums $\%CG_{10}$ and $posDCG_{10}$ for CUI embeddings (cui2vec) are comparable to the maximum values for word embeddings (PubMed and PubMed&PCM). Moreover, the similarity coverage returned by the embedding is generally low: $0.005 \div 0.045$ for $\%CG_{10}$ and $0.007 \div 0.053$ for $posDCG_{10}$, despite both metrics range between 0 and 1. The maximum value for %CG $= 0.045$ means that only 4.5% of the terms returned by a similarity search are in the domain of interest. The ranking of embedding appears different when considers $\%CG_{10}$ and $posDCG_{10}$. For example, among the word embeddings, PubMed&PMC shows the best score for $posDCG_{10}$ whereas PubMed is the best for $\%CG_{10}$. This implies that PubMed returns in average more relevant terms via a similarity search than PubMed&PMC, but PubMed&PMC ranks the relevant terms higher in the returned list of results.

The similarity coverage metrics are more discriminating than the vocabulary coverage: Fig. 1 reveals slight differences among the three best $posDCG_{10}$ ranked embedding, PMC, PubMed, PubMed&PMC, while such differences are not noticeable for $\%IV$.

For the sake of completeness, Fig. 2 shows how $posDCG$ changes, varying k: the performance of $posDCG$ decreases weakly when k increases, as it may be expected from the normalisation factor in eq. (8). The set of experiments related to the vocabulary and similarity coverage are available at the git repo. The order among embeddings slightly changes, and there are no upheavals in the best five embeddings when varying k. Only cui2vec improves its ranking when increasing k. Indeed, considering the low percentage of coverage, selecting the third best-ranked embedding in place of the first or second-ranked might not generally determine a big difference in the performance.

In conclusion, the proposed metrics for similarity coverage provide tools of increasing complexity for inspecting the embeddings and promoting their aware adoption in applications. The poor support of similarity in a specific domain of interest is a warning bell for tuning the overall expectations when adopting the embeddings in particular tasks.

A second part of the experiments applies the analogy measures discussed in section 3.3.2 to evaluate the embeddings for the relations listed in Table 1. All the detailed results are available in Tables 5 and 6, while the experimental pipeline may be found at git repo. Fig. 3, 4 and 5, exemplify the results obtained for Percentage Analogy Coverage (M^{ar}), Analogy Reasoning Coverage (AR), and In-Vocabulary for analogical reasoning (IV^{ar}), conveying the qualitative considerations that follow.

The analysed embeddings support only a limited percentage of the analogies expected by UMLS: the maximum percentage analogy coverage (M^{ar} in Fig. 3) is for the *disease_has_associated_gene* relation is 0.023, which corresponds to 2.3% of the expected seed-constrained UMLS analogical reasonings.

Table 4
Examples of how seeds are covered by the embeddings.

CUI, seed lexical representation	*claims cuis*	*DeVine*	*cui2vec*	*PMC*	*PubMed*	*PubMed&PMC*	*wiki&PubMed&PMC*	*Healthvec*	*GoogleNews*	*tweetsvec*
C0006261, Bronchial Diseases		✓								
C0006264, Bronchial Neoplasms		✓	✓							
C0006266, Bronchospasm		✓	✓	✓	✓	✓	✓	✓	✓	✓
C0006270, Bronchioles		✓	✓	✓	✓	✓	✓		✓	
C0008679, Chronic disease		✓	✓							
C0024109, Lung		✓	✓	✓	✓	✓	✓	✓	✓	✓
C0024115, Lung diseases		✓	✓	✓	✓	✓	✓		✓	
C0024117, Chronic Obstructive Airway Disease		✓	✓	✓	✓	✓	✓		✓	
C4255083, diagnostic imaging aspects				✓	✓	✓	✓		✓	
C0600260, Lung Diseases, Obstructive			✓							
C1969833, COPD, Severe Early-Onset										
C0024121, Lung Neoplasms		✓	✓							
C0034050, Pulmonary Alveolar Proteinosis	✓	✓	✓							
C0206062, Lung Diseases, Interstitial		✓	✓	✓	✓	✓	✓		✓	
C0155883, Chronic Obstructive Asthma	✓		✓							
C0264364, Bronchiolar disease										
C0746102, Chronic lung disease		✓	✓	✓	✓	✓	✓	✓	✓	✓
C0205191, chronic		✓	✓	✓	✓	✓	✓	✓	✓	✓
C0264371, Chronic obliterative bronchiolitis		✓								
C0264220, Chronic disease of respiratory system		✓	✓							

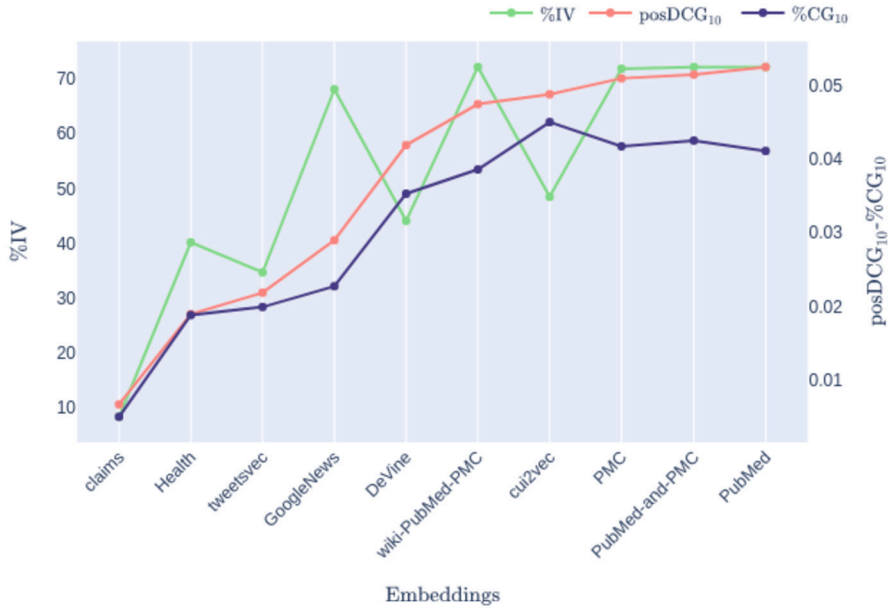

Fig. 1. The figure shows $\%IV$ -in green-, $posDCG_{10}$ -in pink-, $\%CG_{10}$ -in blue- for the analysed embeddings, ordering the models in x-axis by $posDCG_{10}$ performances. The left y-axis shows a percentage scale, for $\%IV$, the right y-axis represents the scale for $\%CG_{10}$ and $posDCG_{10}$. The values refer to $k = 10$, using as domain representation the seed for concepts, obtained making the union between *seeds of UMLS relations* and *seeds by MetaMap*.

The low percentages are due to the normalisation factor $|W_i||W_i - 1|$ in M^{ar} (eq. (16)), which is the overall number of seed-constrained UMLS analogical reasonings for the relationship i. The number of couples that UMLS can form involving seed concepts in the relation i, i.e., $|W_i|$, might be quite big (see Fig. 6), making the number of reasonings to be supported considerably high.

Some embeddings support a remarkable number of analogies, although that number is only a low percentage of those modelled in UMLS. This can be observed considering the non-normalised analogy coverage AR. For example, Fig. 4 shows that cui2vec can close more than 20000 analogies for the relation *finding_site_of*, though 20000 analogies are only a tiny portion of those induced by UMLS (percentage analogy coverage $M^{ar} < 0.005$ in Fig. 3).

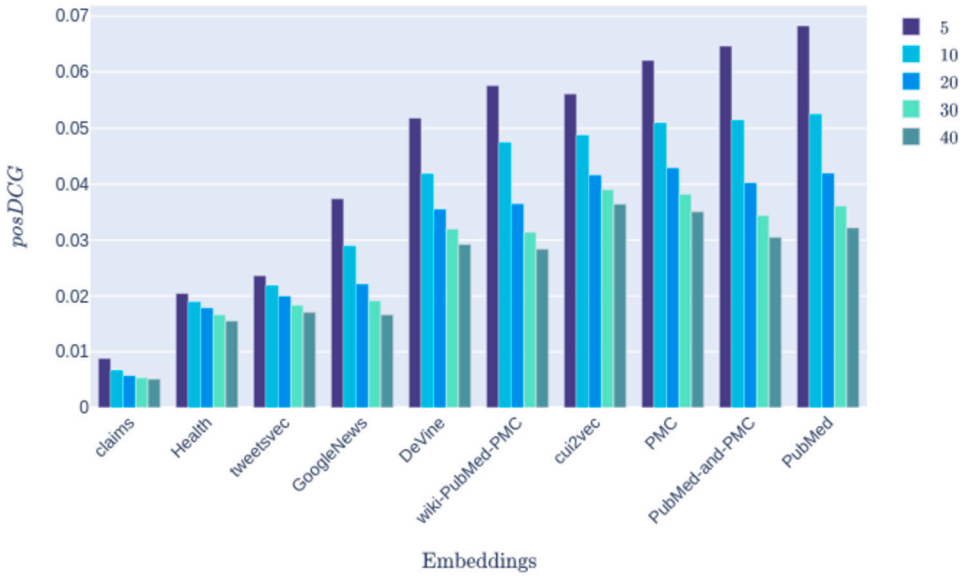

Fig. 2. The values of $posDCG$ varying k: 5,10,20,30,40, correspond to k value for the k-NN. The x axis shows the investigated embeddings sorted by $posDCG_{10}$.

Table 5
Results for Percentage Analogy Coverage (M^{ar}), Analogy Reasoning Coverage (AR) and In-Vocabulary for Analogical Reasoning (IV^{ar}). Selected relationships are indicated only by their initials (e.g., course_of as co, has_course as hc).

	claims_cui			DeVine			cui2vec			Healthvec			tweetsvec		
	M^{ar}	AR	$\frac{\lvert IV^{ar}\rvert}{\lvert W\rvert}$	M^{ar}	AR	$\frac{\lvert IV^{ar}\rvert}{\lvert W\rvert}$	M^{ar}	AR	$\frac{\lvert IV^{ar}\rvert}{\lvert W\rvert}$	M^{ar}	AR	$\frac{\lvert IV^{ar}\rvert}{\lvert W\rvert}$	M^{ar}	AR	$\frac{\lvert IV^{ar}\rvert}{\lvert W\rvert}$
afo	0.00002	1	$\frac{10}{248}$	0.00042	26	$\frac{23}{248}$	0.00318	195	$\frac{58}{248}$	0	0	$\frac{4}{248}$	0	0	$\frac{4}{248}$
haf	0.00008	5		0.00056	34		0.00364	223		0.00002	1		0	0	
amo	0	0	$\frac{24}{2286}$	0.00036	1896	$\frac{527}{2286}$	0.00091	4778	$\frac{583}{2286}$	0	16	$\frac{31}{2286}$	0	1	$\frac{20}{2286}$
ham	0.00001	56		0.00042	2198		0.00043	2237		0.00001	33		0	11	
awmogp	0	0	$\frac{0}{33}$	0	0	$\frac{6}{33}$	0.00284	3	$\frac{14}{33}$	0	0	$\frac{1}{33}$	0	0	$\frac{0}{33}$
gpmawd	0	0		0	0		0.00189	2		0	0		0	0	
cco	0	0	$\frac{0}{1903}$	0	0	$\frac{316}{1903}$	0	5	$\frac{445}{1903}$	0	0	$\frac{11}{1903}$	0	0	$\frac{5}{1903}$
hcc	0	0		0.00003	121		0.00013	451		0	0		0	0	
cwd	0.00049	321	$\frac{122}{811}$	0.00251	1649	$\frac{400}{811}$	0.00284	1863	$\frac{438}{811}$	0.00013	84	$\frac{99}{811}$	0.00002	11	$\frac{44}{811}$
hcd	0.00015	95		0.00181	1190		0.00227	1489		0.00051	337		0	1	
co	0	0	$\frac{0}{1236}$	0.00138	2101	$\frac{315}{1236}$	0.00182	2774	$\frac{409}{1236}$	0	3	$\frac{15}{1236}$	0	0	$\frac{8}{1236}$
hc	0	0		0.00010	150		0.00032	490		0	2		0	2	
dhaas	0	0	$\frac{0}{2387}$	0.00049	2812	$\frac{598}{2387}$	0.00058	3302	$\frac{692}{2387}$	0.00001	43	$\frac{22}{2387}$	0	21	$\frac{15}{2387}$
iaaso	0	0		0.00028	1608		0.00120	6832		0.00001	47		0	13	
dhag	0	0	$\frac{0}{150}$	0	0	$\frac{0}{150}$	0	0	$\frac{1}{150}$	0.00018	4	$\frac{3}{150}$	0	0	$\frac{1}{150}$
gawd	0	0		0	0		0	0		0.00004	1		0	0	
fso	0	0	$\frac{0}{4559}$	0.00023	4728	$\frac{1030}{4559}$	0.00106	21935	$\frac{1281}{4559}$	0.00001	153	$\frac{63}{4559}$	0	42	$\frac{35}{4559}$
hfs	0	0		0.00020	4170		0.00037	7582		0.00001	219		0	36	
mo	0	3	$\frac{21}{1553}$	0.00018	441	$\frac{219}{1553}$	0.00027	642	$\frac{245}{1553}$	0	7	$\frac{9}{1553}$	0	0	$\frac{0}{1553}$
hm	0	1		0.00016	385		0.00056	1358		0	0		0	0	
mt	0.00511	7452	$\frac{240}{1208}$	0.00721	10510	$\frac{563}{1208}$	0.01100	16011	$\frac{664}{1208}$	0.00086	1250	$\frac{163}{1208}$	0.00008	119	$\frac{88}{1208}$
mbtb	0.00228	3326		0.00458	6674		0.00545	7946		0.00284	4141		0.00020	289	

Closing reasonings with inverse relationships does not present a consistent symmetry. That is quite evident comparing In-Vocabulary for analogical reasoning IV^{ar} (Fig. 5) with analogical reasoning AR (Fig. 4). In-Vocabulary for analogical reasoning IV^{ar} (Fig. 5) shows the number of couples that can be found in the embedding, but does not measure if the relation i is actually represented. In fact, Fig. 5 shows the same values for the properties that are the inverse of each other (e.g., *may_treat* and *may_be_treated_by* or *associated_finding_of* and *has_associated_finding*). On the contrary, analogical reasoning AR represents the number of analogies an embedding closes via relation i. Columns for inverse relations score mostly different for embeddings in Fig. 4, since analogies closed for a relation are not necessarily closed for its inverse relation.

The proposed analogy coverage measures provide tools to inspect the embeddings. Whereas AR represents a mere count of analogies, M^{ar} represents a percentage of UMLS-derived analogies. The former could return high values that, when weighted, might appear negligible. Vice versa, relatively low scores of AR could correspond to high values of M^{ar}.

Which embedding is the best might depend on the specific goals and relations considered, as embeddings do not support all the relations equally. For example, PubMed covers the highest percentage of UMLS-driven reasoning for the relationship *disease_has_as-*

Table 6
Results for Percentage Analogy Coverage (M^{ar}), Analogy Reasoning Coverage (AR) and In-Vocabulary for Analogical Reasoning (IV^{ar}). Selected relationships are indicated only by their initials (e.g., course_of as co, has_course as hc).

	PMC			PubMed			PubMed&PMC			wiki&PubMed&PMC			GoogleNews		
	M^{ar}	AR	$\frac{\lvert IV^{ar}\rvert}{\lvert W\rvert}$	M^{ar}	AR	$\frac{\lvert IV^{ar}\rvert}{\lvert W\rvert}$	M^{ar}	AR	$\frac{\lvert IV^{ar}\rvert}{\lvert W\rvert}$	M^{ar}	AR	$\frac{\lvert IV^{ar}\rvert}{\lvert W\rvert}$	M^{ar}	AR	$\frac{\lvert IV^{ar}\rvert}{\lvert W\rvert}$
afo	0	0	$\frac{9}{248}$	0	0	$\frac{9}{248}$	0.00002	1	$\frac{9}{248}$	0.00003	2	$\frac{9}{248}$	0	0	$\frac{8}{248}$
haf	0.00003	2	$\frac{9}{248}$	0	0	$\frac{9}{248}$	0	0	$\frac{9}{248}$	0	0	$\frac{9}{248}$	0	0	$\frac{8}{248}$
amo	0.00004	189	$\frac{97}{2286}$	0.00005	241	$\frac{105}{2286}$	0.00007	364	$\frac{105}{2286}$	0.00004	222	$\frac{106}{2286}$	0.00002	95	$\frac{78}{2286}$
ham	0.00001	36	$\frac{97}{2286}$	0.00001	69	$\frac{105}{2286}$	0.00001	63	$\frac{105}{2286}$	0.00001	54	$\frac{106}{2286}$	0	8	$\frac{78}{2286}$
awmogp	0.01420	15	$\frac{27}{33}$	0.01800	19	$\frac{27}{33}$	0.01800	19	$\frac{27}{33}$	0.01610	17	$\frac{27}{33}$	0.01230	13	$\frac{19}{33}$
gpmawd	0.00379	4	$\frac{27}{33}$	0.00095	1	$\frac{27}{33}$	0.00189	2	$\frac{27}{33}$	0	0	$\frac{27}{33}$	0.00095	1	$\frac{19}{33}$
cco	0	0	$\frac{71}{1903}$	0	0	$\frac{73}{1903}$	0	0	$\frac{75}{1903}$	0	0	$\frac{75}{1903}$	0	0	$\frac{46}{1903}$
hcc	0.00001	20	$\frac{71}{1903}$	0.00001	27	$\frac{73}{1903}$	0.00001	36	$\frac{75}{1903}$	0.00001	36	$\frac{75}{1903}$	0	3	$\frac{46}{1903}$
cwd	0.00054	357	$\frac{232}{811}$	0.00092	607	$\frac{241}{811}$	0.00086	575	$\frac{241}{811}$	0.00085	561	$\frac{241}{811}$	0.00014	91	$\frac{163}{811}$
hcd	0.00222	1460	$\frac{232}{811}$	0.00370	2426	$\frac{241}{811}$	0.00165	1086	$\frac{241}{811}$	0.00203	1331	$\frac{241}{811}$	0.00316	2074	$\frac{163}{811}$
co	0.00016	242	$\frac{68}{1236}$	0.00011	172	$\frac{71}{1236}$	0.00016	245	$\frac{72}{1236}$	0.00015	228	$\frac{72}{1236}$	0.00003	52	$\frac{47}{1236}$
hc	0.00001	18	$\frac{68}{1236}$	0.00002	25	$\frac{71}{1236}$	0.00002	27	$\frac{72}{1236}$	0.00002	29	$\frac{72}{1236}$	0	2	$\frac{47}{1236}$
dhaas	0.00001	65	$\frac{78}{2387}$	0.00002	100	$\frac{81}{2387}$	0.00002	102	$\frac{81}{2387}$	0.00001	80	$\frac{81}{2387}$	0.00001	39	$\frac{66}{2387}$
iaaso	0.00004	226	$\frac{78}{2387}$	0.00004	208	$\frac{81}{2387}$	0.00005	270	$\frac{81}{2387}$	0.00004	250	$\frac{81}{2387}$	0.00001	56	$\frac{66}{2387}$
dhag	0.01650	368	$\frac{105}{150}$	0.02300	514	$\frac{100}{150}$	0.01610	359	$\frac{108}{150}$	0.01510	338	$\frac{108}{150}$	0.00966	216	$\frac{71}{150}$
gawd	0.00107	24	$\frac{105}{150}$	0.00107	24	$\frac{100}{150}$	0.00094	21	$\frac{108}{150}$	0.00103	23	$\frac{108}{150}$	0.00018	4	$\frac{71}{150}$
fso	0.00003	584	$\frac{214}{4559}$	0.00004	782	$\frac{229}{4559}$	0.00004	943	$\frac{230}{4559}$	0.00003	724	$\frac{232}{4559}$	0.00001	235	$\frac{166}{4559}$
hfs	0.00001	169	$\frac{214}{4559}$	0.00001	294	$\frac{229}{4559}$	0.00001	217	$\frac{230}{4559}$	0.00001	208	$\frac{232}{4559}$	0.00001	117	$\frac{166}{4559}$
mo	0.00097	2353	$\frac{409}{1553}$	0.00183	4417	$\frac{447}{1553}$	0.00185	4468	$\frac{447}{1553}$	0.00196	4726	$\frac{490}{1553}$	0.00052	1254	$\frac{268}{1553}$
hm	0.00008	198	$\frac{409}{1553}$	0.00012	280	$\frac{447}{1553}$	0.00007	171	$\frac{447}{1553}$	0.00013	311	$\frac{490}{1553}$	0	8	$\frac{268}{1553}$
mt	0.00189	2761	$\frac{391}{1208}$	0.00253	3684	$\frac{426}{1208}$	0.00252	3677	$\frac{427}{1208}$	0.00259	3781	$\frac{429}{1208}$	0.00046	671	$\frac{265}{1208}$
mbtb	0.00325	4741	$\frac{391}{1208}$	0.01020	14847	$\frac{426}{1208}$	0.00543	7916	$\frac{427}{1208}$	0.00670	9775	$\frac{429}{1208}$	0.00252	3675	$\frac{265}{1208}$

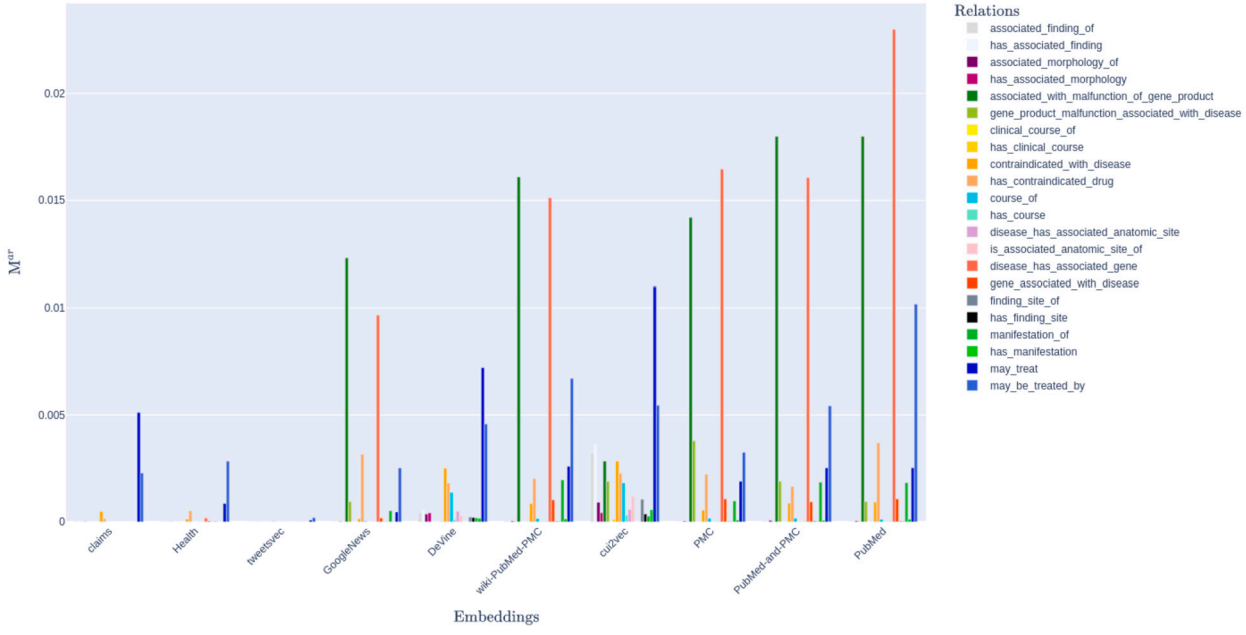

Fig. 3. The Percentage Analogy Coverage M_i^{ar} for a given relation i.

sociated_gene, but cui2vec outperforms PubMed when dealing with *may_treat*, see Fig. 3. If the aim is not supporting all the possible UMLS-driven reasonings, but rather enabling analogy for the widest number of couples, ones might privilege the AR measure in place of M^{ar}. According to the AR measure, cui2vec shows generally better performances. Even if DeVine has lower performances than cui2vec in terms of AR, it is the only embeddings except cui2vec showing $M^{ar} > 0$ for *associated_finding_of* and its inverse, *associated_morphology_of* and its inverse, and *course_of* and its inverse.

In conclusion, the analogy measures assess different aspects of domain coverage by the embeddings: IV_i^{ar} quantifies the number of contemplated *potential couples* linked by the relationship i, assuring the presence of the couple elements inside the embedding vocabulary; AR_i measures the actual number of solved permutations between the aforesaid couples, whose result is weighted with the partial permutation of *all the couples* in UMLS for the relationship i in M_i^{ar}; AR_i, as already described, provides a very detailed picture of the relationship inside the considered embedding, but, as side effect, it requires very high computational cost, particularly

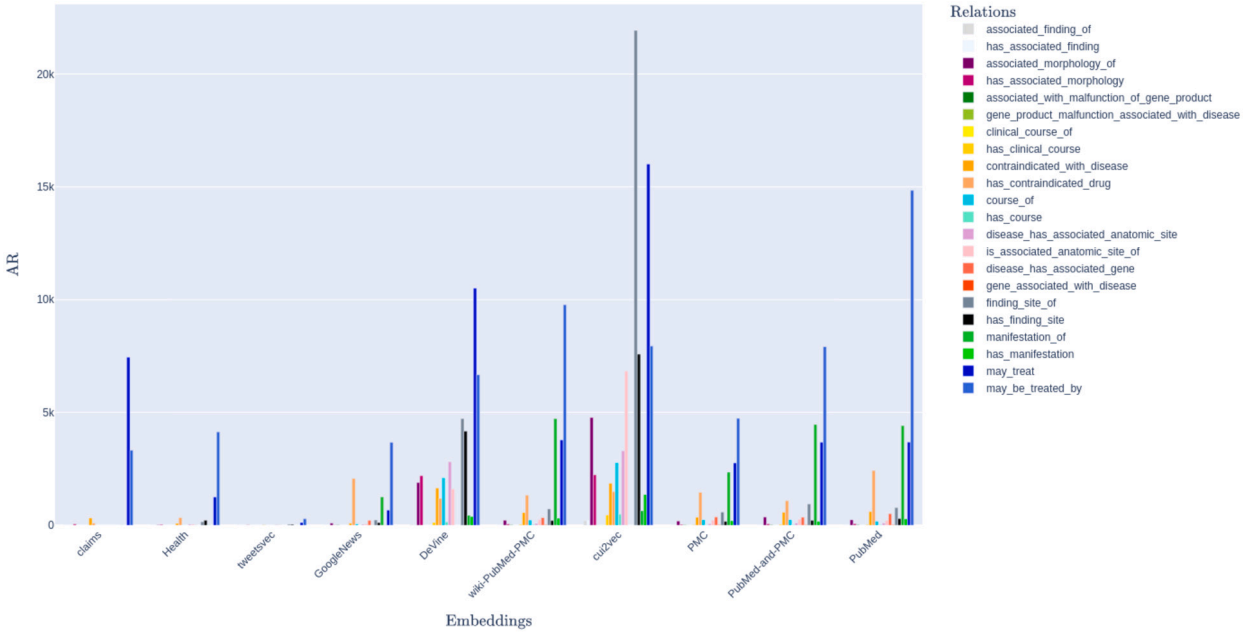

Fig. 4. The Analogy Reasoning Coverage AR_i for a given relation i.

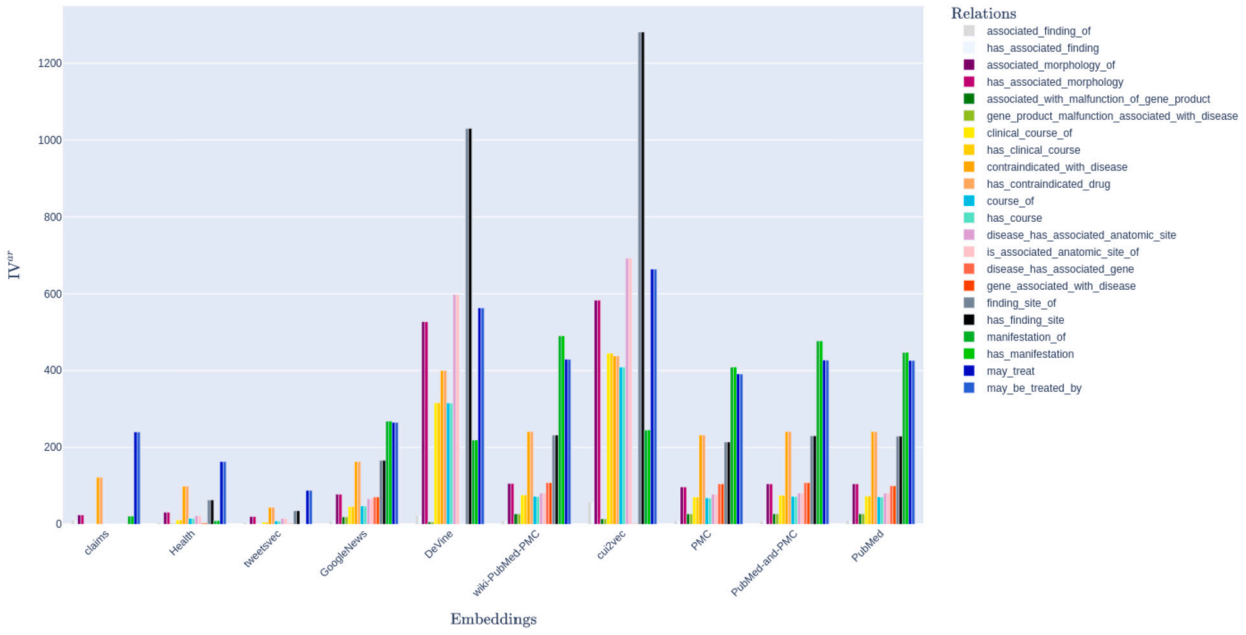

Fig. 5. The In-Vocabulary for analogy reasoning IV_i^{ar} for a given relation i.

for benchmarking among several embeddings. This issue is partially overcome by the variety of the proposed measures: the "lightest" IV_i^{ar} would allow an initial sorting of the resources, hence for more accurate benchmarking AR_i and M_i^{ar} are suggested.

5. Conclusions

In this paper we introduced a methodology to evaluate the quality of biomedical embeddings through bespoke measures of domain coverage. We targetted the specific branch of chronic obstructive pulmonary disease not only addressing the terminological coverage, but also how well the embedding supports similarities, relatedness and analogical reasoning within the domain. To the best of our knowledge, we have been the first to deal with this definition of coverage for a specific domain and we believe this approach is useful to select the most appropriate embedding in real use scenarios. In particular, we tailored existing measures and run the

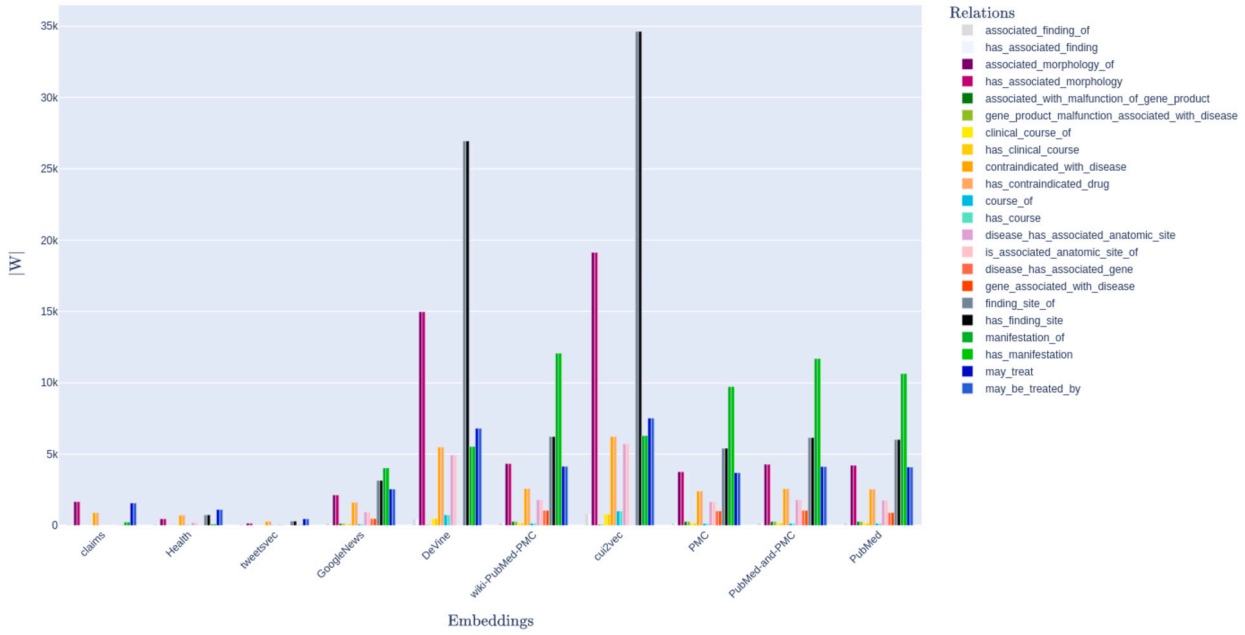

Fig. 6. The cardinality of the seeds for analogy $|W_i|$ for a given relation i.

experimentation on several available embeddings, spanning from unstructured to structured ones, from word to CUI ones. We may conclude that the coverage on the COPD subdomain is generally poor, referring to the recent literature. Similarly, embeddings may misrepresent other specialised application areas, and our methodology can be deployed for testing the quality of models in these domains. In the future, we plan to face this issue, considering different biomedical domains, e.g., neurological disorders.

The general motivation of this paper resides in the issue of the best reuse of resources, which became a crucial point with the complexity and modularisation of the current systems. Indeed, intrinsic model evaluation is more and more important to select and exploit the most suitable available resources independently of specific downstream tasks: according to the principle of trustworthy AI, model quality needs to be documented in order to guarantee a fair adoption of technologies.

In this article we considered only static embeddings, which still have several advantages and may be interpreted and used as an alternative knowledge representation tool. Nevertheless, contextual embeddings have been gaining more and more importance thanks to their flexibility and potential in many applications, including the biomedical field, and consequently are worth to be explored. Among those, we shall mention BioBERT [21] and derivatives. Methods for the performance evaluation of BERT-family models in the biomedical field are present in literature [32]. Such methods are based on the characteristic fine tuning of BERT, which is more similar to extrinsic evaluation. We could not find measurable methodologies as we proposed in this paper. It would be possible to convert contextual embeddings in static ones and apply our approach, but the cost of such conversion is too high to make the problem tractable. Contextual embeddings will be then tackled in a future work.

CRediT authorship contribution statement

Salvatore Giancani: Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Riccardo Albertoni, Chiara Eva Catalano: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Salvatore Giancani reports financial support was provided by “Advances in pneumology via ICT and data analytics” (PNEULYTICS) funded by Compagnia di San Paolo (Scientific call 2019), ID ROL: 34754. Riccardo Albertoni reports financial support was provided by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215. Chiara Eva Catalano reports financial support was provided by “Advances in pneumology via ICT and data analytics” (PNEULYTICS) funded by Compagnia di San Paolo (Scientific call 2019), ID ROL: 34754. Chiara Eva Catalano reports financial support was provided by the project “RAISE - Robotics and AI for Socio-economic Empowerment”, supported by European Union - NextGenerationEU.

Data availability statement

Data associated with this study has been deposited at https://github.com/SaGiancani/medical-concepts-embeddings.

Acknowledgements

This research was partially supported by the project “Advances in pneumology via ICT and data analytics” (PNEULYTICS) funded by Compagnia di San Paolo (Scientific call 2019), ID ROL: 34754. It was also partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215, and was partially carried out within the framework of the project “RAISE - Robotics and AI for Socio-economic Empowerment” and has been supported by European Union - NextGenerationEU.

References

[1] F. Alshargi, S. Shekarpour, T. Soru, A.P. Sheth, Metrics for evaluating quality of embeddings for ontological concepts, in: A. Martin, K. Hinkelmann, A. Gerber, D. Lenat, F. van Harmelen, P. Clark (Eds.), Proceedings of the AAAI 2019 Spring Symposium on Combining Machine Learning with Knowledge Engineering (AAAI-MAKE 2019), Stanford University, Palo Alto, California, USA, March 25–27, 2019, in: CEUR Workshop Proceedings, vol. 2350, 2019, CEUR-WS.org, http://ceur-ws.org/Vol-2350/paper26.pdf.

[2] A.R. Aronson, Effective mapping of biomedical text to the umls metathesaurus: the metamap program, in: Proceedings of the AMIA Symposium, American Medical Informatics Association, 2001, p. 17.

[3] A.L. Beam, B. Kompa, A. Schmaltz, I. Fried, G. Weber, N. Palmer, X. Shi, T. Cai, I.S. Kohane, Clinical concept embeddings learned from massive sources of multimodal medical data, in: Pacific Symposium on Biocomputing 2020, World Scientific, 2019, pp. 295–306.

[4] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic Acids Res. 32 (suppl_1) (2004) D267–D270.

[5] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Trans. Assoc. Comput. Linguist. (ISSN 2307-387X) 5 (2017) 135, https://doi.org/10.1162/tacl_a_00051.

[6] R. Bommasani, K. Davis, C. Cardie, Interpreting pretrained contextualized representations via reductions to static embeddings, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4758–4781.

[7] A. Callahan, J. Cruz-Toledo, P. Ansell, M. Dumontier, Bio2rdf release 2: improved coverage, interoperability and provenance of life science linked data, in: P. Cimiano, O. Corcho, V. Presutti, L. Hollink, S. Rudolph (Eds.), The Semantic Web: Semantics and Big Data, Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN 978-3-642-38288-8, 2013, pp. 200–212.

[8] B.R. Celli, W. MacNee, A. Agusti, A. Anzueto, B. Berg, A.S. Buist, P.M. Calverley, N. Chavannes, T. Dillard, B. Fahy, et al., Standards for the diagnosis and treatment of patients with copd: a summary of the ats/ers position paper, Eur. Respir. J. 23 (6) (2004) 932–946.

[9] 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 (Suppl 2) (2018) 12911, https://doi.org/10.1186/s12911-018-0630-x (ISSN 14726947).

[10] B. Chiu, S. Baker, Word embeddings for biomedical natural language processing: a survey, Lang. Linguist. Compass 14 (12) (2020), https://doi.org/10.1111/lnc3.12402.

[11] Y. Choi, C.Y.-I. Chiu, D. Sontag, Learning low-dimensional representations of medical concepts, AMIA Summits Transl. Sci. Proc. 2016 (2016) 41.

[12] F. Dassereto, L. Di Rocco, G. Guerrini, M. Bertolotto, Evaluating the effectiveness of embeddings in representing the structure of geospatial ontologies, in: P. Kyriakidis, D. Hadjimitsis, D. Skarlatos, A. Mansourian (Eds.), Geospatial Technologies for Local and Regional Development, Springer International Publishing, Cham, ISBN 978-3-030-14745-7, 2020, pp. 41–57.

[13] L. De Vine, G. Zuccon, B. Koopman, L. Sitbon, P. Bruza, Medical semantic similarity with a neural language model, in: Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, 2014, pp. 1819–1822.

[14] European Commission, Directorate General for Communications Networks, Content and Technology, The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self Assessment, Publications Office, LU, 2020, https://data.europa.eu/doi/10.2759/002360.

[15] M. Faruqui, Y. Tsvetkov, P. Rastogi, C. Dyer, Problems with evaluation of word embeddings using word similarity tasks, in: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, Berlin, Germany, Association for Computational Linguistics, Aug. 2016, pp. 30–35, https://doi.org/10.18653/v1/W16-2506, https://aclanthology.org/W16-2506.

[16] J.R. Firth, A synopsis of linguistic theory, 1930–1955, in: Studies in Linguistic Analysis, 1957.

[17] M. Günther, P. Sikorski, M. Thiele, W. Lehner. Facete, Exploiting web tables for domain-specific word embedding evaluation, in: Proceedings of the Workshop on Testing Database Systems, DBTest ’20, New York, NY, USA, Association for Computing Machinery, ISBN 9781450380010, 2020, https://doi.org/10.1145/3395032.3395325.

[18] P. Gärdenfors, Conceptual Spaces: The Geometry of Thought, The MIT Press, ISBN 9780262273558, 03 2000, https://doi.org/10.7551/mitpress/2076.001.0001.

[19] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of ir techniques, ACM Trans. Inf. Syst. 20 (4) (2002) 422–446.

[20] K.S. Kalyan, S. Sangeetha, Secnlp: a survey of embeddings in clinical natural language processing, J. Biomed. Inform. 101 (2020) 103323.

[21] 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, Bioinformatics 36 (4) (2020) 1234–1240.

[22] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P.N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, C. Bizer, Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia, Semant. Web 6 (2) (2015) 167–195.

[23] O. Levy, Y. Goldberg, Dependency-based word embeddings, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Stroudsburg, PA, USA, Association for Computational Linguistics, ISBN 9781420016789, 2014, pp. 302–308, https://doi.org/10.3115/v1/P14-2050, http://aclweb.org/anthology/P14-2050.

[24] O. Levy, Y. Goldberg, Linguistic regularities in sparse and explicit word representations, in: Proceedings of the Eighteenth Conference on Computational Natural Language Learning, Ann Arbor, Michigan, Association for Computational Linguistics, June 2014, pp. 171–180, https://doi.org/10.3115/v1/W14-1618, https://aclanthology.org/W14-1618.

[25] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013.

[26] S. Moen, T.S.S. Ananiadou, Distributional semantics resources for biomedical text processing, in: Proceedings of LBM, 2013, pp. 39–44.

[27] A. Nikfarjam, A. Sarker, K. O’connor, R. Ginn, G. Gonzalez, Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features, J. Am. Med. Inform. Assoc. 22 (3) (2015) 671–681.

[28] J. Noh, R. Kavuluru, Improved biomedical word embeddings in the transformer era, J. Biomed. Inform. 120 (2021) 103867.

[29] S. Pakhomov, B. McInnes, T. Adam, Y. Liu, T. Pedersen, G.B. Melton, Semantic similarity and relatedness between clinical terms: an experimental study, in: AMIA Annual Symposium Proceedings, vol. 2010, American Medical Informatics Association, 2010, p. 572.

[30] S.V. Pakhomov, T. Pedersen, B. McInnes, G.B. Melton, A. Ruggieri, C.G. Chute, Towards a framework for developing semantic relatedness reference standards, J. Biomed. Inform. 44 (2) (apr 2011) 251–265, https://doi.org/10.1016/j.jbi.2010.10.004 (ISSN 15320464), https://pubmed.ncbi.nlm.nih.gov/21044697/.

[31] T. Pedersen, S.V. Pakhomov, S. Patwardhan, C.G. Chute, Measures of semantic similarity and relatedness in the biomedical domain, J. Biomed. Inform. 40 (3) (2007) 288–299.
[32] Y. Peng, S. Yan, Z. Lu, Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets, arXiv preprint, arXiv:1906.05474, 2019.
[33] M. Samwald, A. Jentzsch, C. Bouton, C. Kallesøe, E.L. Willighagen, J.G. Hajagos, M.S. Marshall, E. Prud'hommeaux, O. Hassanzadeh, E. Pichler, S. Stephens, Linked open drug data for pharmaceutical research and development, J. Cheminform. 3 (19) (2011).
[34] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, C.G. Chute, Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. 17 (5) (2010) 507–513.
[35] T. Schnabel, I. Labutov, D. Mimno, T. Joachims, Evaluation methods for unsupervised word embeddings, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, Association for Computational Linguistics, Sept. 2015, pp. 298–307, https://doi.org/10.18653/v1/D15-1036, https://aclanthology.org/D15-1036.
[36] M.Z. Sh, E. Tutubalina, A. Tropsha, Identifying disease-related expressions in reviews using conditional random fields, Komp'jut. Lingvistika Intellekt. Tehnologii 1 (16) (2017) 155–166.
[37] G. Soğancıoğlu, H. Öztürk, A. Özgür, Biosses: a semantic sentence similarity estimation system for the biomedical domain, Bioinformatics 33 (14) (2017) i49–i58.
[38] H. Turki, T. Shafee, M.A.H. Taieb, M.B. Aouicha, D. Vrandecic, D. Das, H. Hamdi Wikidata, A large-scale collaborative ontological medical database, J. Biomed. Inform. 99 (2019).
[39] B. Wang, A. Wang, F. Chen, Y. Wang, C.-C.J. Kuo, Evaluating word embedding models: methods and experimental results, APSIPA Trans. Signal Inf. Process. 8 (2019) e19, https://doi.org/10.1017/ATSIP.2019.12.
[40] H. Wang, H. Du, G. Qi, H. Chen, W. Hu, Z. Chen, et al., Construction of a linked data set of covid-19 knowledge graphs: development and applications, JMIR Med. Inform. 10 (5) (2022) e37215.
[41] Y. Wang, L. Wang, Y. Li, D. He, W. Chen, T.-Y. Liu, A theoretical analysis of ndcg ranking measures, in: Proceedings of the 26th Annual Conference on Learning Theory (COLT 2013), vol. 8, Citeseer, 2013, p. 6.
[42] 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 (July 2018) 12–20, https://doi.org/10.1016/j.jbi.2018.09.008 (ISSN 15320464).
[43] Y. Wang, N. Afzal, S. Fu, L. Wang, F. Shen, M. Rastegar-Mojarad, H. Liu, Medsts: a resource for clinical semantic textual similarity, Lang. Resour. Eval. 54 (1) (2020) 57–72.