


DATABASE

Open Access



BioVerbNet: a large semantic-syntactic classification of verbs in biomedicine

Olga Majewska^{1*} , Charlotte Collins¹, Simon Baker¹, Jari Björne², Susan Windisch Brown³, Anna Korhonen¹ and Martha Palmer³

Abstract

Background: Recent advances in representation learning have enabled large strides in natural language understanding; However, verbal reasoning remains a challenge for state-of-the-art systems. External sources of structured, expert-curated verb-related knowledge have been shown to boost model performance in different Natural Language Processing (NLP) tasks where accurate handling of verb meaning and behaviour is critical. The costliness and time required for manual lexicon construction has been a major obstacle to porting the benefits of such resources to NLP in specialised domains, such as biomedicine. To address this issue, we combine a neural classification method with expert annotation to create BioVerbNet. This new resource comprises 693 verbs assigned to 22 top-level and 117 fine-grained semantic-syntactic verb classes. We make this resource available complete with semantic roles and VerbNet-style syntactic frames.

Results: We demonstrate the utility of the new resource in boosting model performance in document- and sentence-level classification in biomedicine. We apply an established retrofitting method to harness the verb class membership knowledge from BioVerbNet and transform a pretrained word embedding space by pulling together verbs belonging to the same semantic-syntactic class. The BioVerbNet knowledge-aware embeddings surpass the non-specialised baseline by a significant margin on both tasks.

Conclusion: This work introduces the first large, annotated semantic-syntactic classification of biomedical verbs, providing a detailed account of the annotation process, the key differences in verb behaviour between the general and biomedical domain, and the design choices made to accurately capture the meaning and properties of verbs used in biomedical texts. The demonstrated benefits of leveraging BioVerbNet in text classification suggest the resource could help systems better tackle challenging NLP tasks in biomedicine.

Keywords: Verb lexicon, VerbNet, Text classification

Background

The demand for automatic systems capable of processing and mining the rapidly expanding body of biomedical literature is constantly growing and NLP technologies can play a key role in facilitating the dissemination and consolidation of knowledge recorded in scientific papers, patient reports, or clinical notes. The domain-specific properties

of biomedical texts require specialised systems sensitive to the well-defined semantics and syntactic behaviour of the terms used in the scientific literature. This is why high-quality, rich computational lexicons comprising information about the meaning and combinatorial properties of words in biomedical texts can significantly boost the performance of NLP systems in problems ranging from information retrieval, relation and event extraction, or entailment detection. Similarly to the general language domain, lexicographic efforts in biomedicine have primarily focused on nouns (e.g., UMLS Metathesaurus [1]),

*Correspondence: om304@cam.ac.uk

¹Language Technology Laboratory, MMLL, University of Cambridge, 9 West Road, CB39DB Cambridge, UK

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

while the demand for rich, large-coverage verb-specific biomedical resources has not yet been satisfied [2–6].

A number of works in general domain NLP have illustrated the benefits offered by databases of structured verb-related knowledge. One such resource is VerbNet [7], a broad-coverage, hierarchical classification of English verbs providing detailed documentation of verbs' semantic and syntactic properties. It has been successfully employed to boost NLP tasks such as word sense disambiguation [8], semantic role labelling [9], information extraction [10], or text mining [11, 12]. While the utility of VerbNet in the general domain has been widely recognised, the lexicographic effort involved in its construction poses a challenge to transferring its benefits to other specialised domains.

In this work, we address the demand for a biomedical verb resource and alleviate the issue of slow manual dataset construction by combining a data-driven automatic classification approach with post-hoc expert verification and annotation to create the first BioVerbNet. We take the output of a highly accurate neural classification approach of Chiu et al. [13] as a starting point and subsequently manually validate the resultant classes based on VerbNet class criteria of semantic-syntactic coherence of member verbs. The 22 top-level and 117 fine-grained verb classes produced in the process are then annotated with semantic roles of the verbs' arguments and syntactic frames in which class members participate, thus yielding a rich semantic-syntactic lexicon of biomedical verbs.

The creation of BioVerbNet involved the following key stages. First, the 1149 verbs assigned to 50 classes by the system of Chiu et al. were reviewed by a domain expert and a linguist, to identify noisy candidates to be eliminated from the classification. Next, each class was individually validated by verifying each individual candidate member's consistency with the rest of the class in terms of closeness of meaning and shared structural properties, based on the most frequent dependency contexts extracted from the PubMed corpus [14] by Chiu et al. In the process, the experts decided whether misclassified candidates should be (a) reassigned to another existing class, (b) assigned to a new class, or (c) discarded from the classification. We examined the domain specificity of our classification by comparing BioVerbNet to VerbNet, which revealed a very limited coverage of biomedical verbs in VerbNet and important discrepancies in the dominant senses represented by the verbs shared by both resources. Next, for each class a set of representative syntactic contexts was selected, each subsequently annotated with syntactic descriptions, capturing the possible surface realisations of the member verbs' arguments, and their semantic roles. In order to better capture the characteristic properties of the entities acting as Agents in biomedical scenarios (e.g., cells, chemical reactions,

biological processes), we introduced a new biomedicine-specific role of *Bio-Agent*, distinct from the canonical agentive arguments (e.g., human actors) in the general language domain.

We demonstrate the utility of the newly created verb lexicon to support neural approaches to two biomedical text classification tasks. We derive verb class knowledge in the form of pairwise constraints extracted from the BioVerbNet classification, which we employ to retrofit the vector space of pretrained word embeddings to better reflect the shared semantic-syntactic properties represented by each verb class by pulling co-members closer together. We input the BioVerbNet-specialised embeddings into a convolutional neural network model and evaluate it on document- and sentence-level classification tasks using two established biomedical datasets, the Hallmarks of Cancer [15] and the Exposure taxonomy [16]. The promising results achieved by the model boosted with BioVerbNet class information holds promise for future applications of the resource in downstream tasks in biomedicine.

Related work

Computational verb resources

English-language general domain NLP has a number of large-scale expert-built resources at its disposal from which to derive rich information about verb behaviour. These include, among others, the large semantic network WordNet [17], FrameNet [18], which organises concepts in the so-called semantic frames, describing different types of events, relations, entities, and their participants, and PropBank [19], which includes information about semantic propositions and predicate-argument structure of verbal predicates. A lexicon focused exclusively on verbs is VerbNet, which extends Levin's [20] taxonomy of English verbs and groups them into classes based on shared semantic-syntactic properties. Each such class is accompanied by a set of frames, including a syntactic description and a semantic representation, as well as thematic roles and selectional restrictions on the verbs' arguments. VerbNet classes capture useful generalisations about verb behaviour and can boost NLP systems' predictive capacity on unseen vocabulary by providing a means of extrapolating from individual word types to classes. For instance, by linking an unseen verb *quell* to its class SUBJUGATE, a system can refine its meaning representation to align more closely with other, seen class members with higher occurrence rates in corpora (e.g., *suppress*, *dampen*). VerbNet has been used as a source of syntactic and semantic features supporting a range of NLP applications, including machine translation [21], semantic parsing [22], word sense disambiguation [8, 23], information extraction [10] and text mining [11, 12]. While VerbNet offers vast coverage (it currently includes 9344 verbs organised in 329 main classes), its utility cannot

be directly extended to specialised domains, such as biomedicine, where verbs occur in domain-specific senses and distinct contexts, different from their patterns of behaviour in general English. This is why creation of resources tailored to the characteristics of biomedical texts and terminology is essential.

Biomedical lexicons

A large biomedical lexical resource available is the UMLS Metathesaurus, the most-extensive thesaurus in this domain, which classifies concepts pertaining to biomedicine by semantic type and stores information about the relationships among them. While the resource has been used to support biomedical data mining and information retrieval, its focus is on nouns. The currently available verb-specific lexicons have much smaller coverage and are usually limited to narrow sub-domains. For instance, the manually-created UMLS SPECIALIST lexicon is focused on medical and health-related terminology, whereas the BioLexicon, which provides syntactic and semantic frame information for biomedical verbs, is extracted from the *Escherichia Coli* (*E. Coli*) corpora, which restricts its utility to this particular subdomain.

Representation learning and text classification in BioNLP

With deep learning, representation learning has become a standard technique in natural language processing and also in biomedical natural language processing. The utilization of representation learning methods in BioNLP has followed the introduction of such methods in general NLP, and has usually been accompanied by efforts to adjust and specialize these methods for a biomedical vocabulary. In the development of BioNLP research during the past decade, the introduction of Word2Vec [24] has prepared the way for wider use of neural concept representations [25].

The most common use case for word vectors has been as input embeddings for deep learning neural networks, but word vectors have also been used directly for analysing concepts such as semantic similarity and relatedness [26]. Word vectors trained on general domain texts such as news articles may not always have captured the specific semantics of biomedical concepts, so the methods of Word2Vec, GloVe [27] and FastText [28] have been adapted for the generation of specialized vector space representations usually based on the PubMed collection of millions of biomedical research articles [29–31]. In the BioWordVec project [32] further information from MeSH (Medical Subject Headings) is used to augment PubMed text resources. Wang et al. [33] have shown that training embeddings specifically on biomedical text can produce more relevant vector space representations.

The introduction of generalized language models like ELMo and BERT [34, 35] with their integrated embedding

vocabularies introduced a new, more unified approach for utilization of representation learning in language models. As with the word vector models, ELMo and BERT were also rapidly adapted for the specifics of biomedical language [36, 37].

With the advent of deep learning, representation learning has become a common technique in many text mining tasks such as classification. Neural networks based on convolutional and recurrent (especially LSTM) approaches have achieved significantly improved results on many biomedical text mining tasks [38–40].

Transformer models have resulted in even larger performance gains [25]. The BioBERT model itself demonstrated state of the art performance on named entity recognition, relation extraction and question answering. Since its publication this model has been applied to tasks such as drug–drug interaction extraction, classification of social media health discussions and analysis of scientific articles related to COVID-19 [41–43].

Construction and content

Dataset design

The starting point for the construction of BioVerbNet is the automatic classification produced by Chiu et al. [13], which consists of 1149 verbs assigned to 50 classes. Their qualitative evaluation of a small subset of the resource showed that the classes were highly accurate. In this work, we perform a complete manual verification and restructuring of the automatically generated classification, which produces a new, two-level taxonomy of verbs, including 22 top-level classes and 117 subclasses, illustrated in Fig. 1. Next, we carry out two stages of manual annotation, which yield VerbNet-style semantic-syntactic classes, each described by a set of syntactic frames annotated with semantic roles. In the next sections, we first describe the methodology of Chiu et al. [13] and the resultant automatically generated classes, followed by the process of manual verification.

Automatic verb classification

Chiu et al. [13] proposed an automatic classification method which combines a neural representation learning approach with the classification step. In contrast to previous approaches to automatic induction of verb classes [44–47], the method avoids manual feature engineering, which is time-consuming and requires expert knowledge. Instead, it employs features automatically learned directly from corpora using neural networks, fine-tuned to better capture the semantics of verbs in biomedical texts. Due to the cost-effectiveness of the approach and its demonstrated success in generating high-quality classification output, as validated by human experts in Chiu et al., we chose to leverage its potential in our construction of BioVerbNet.

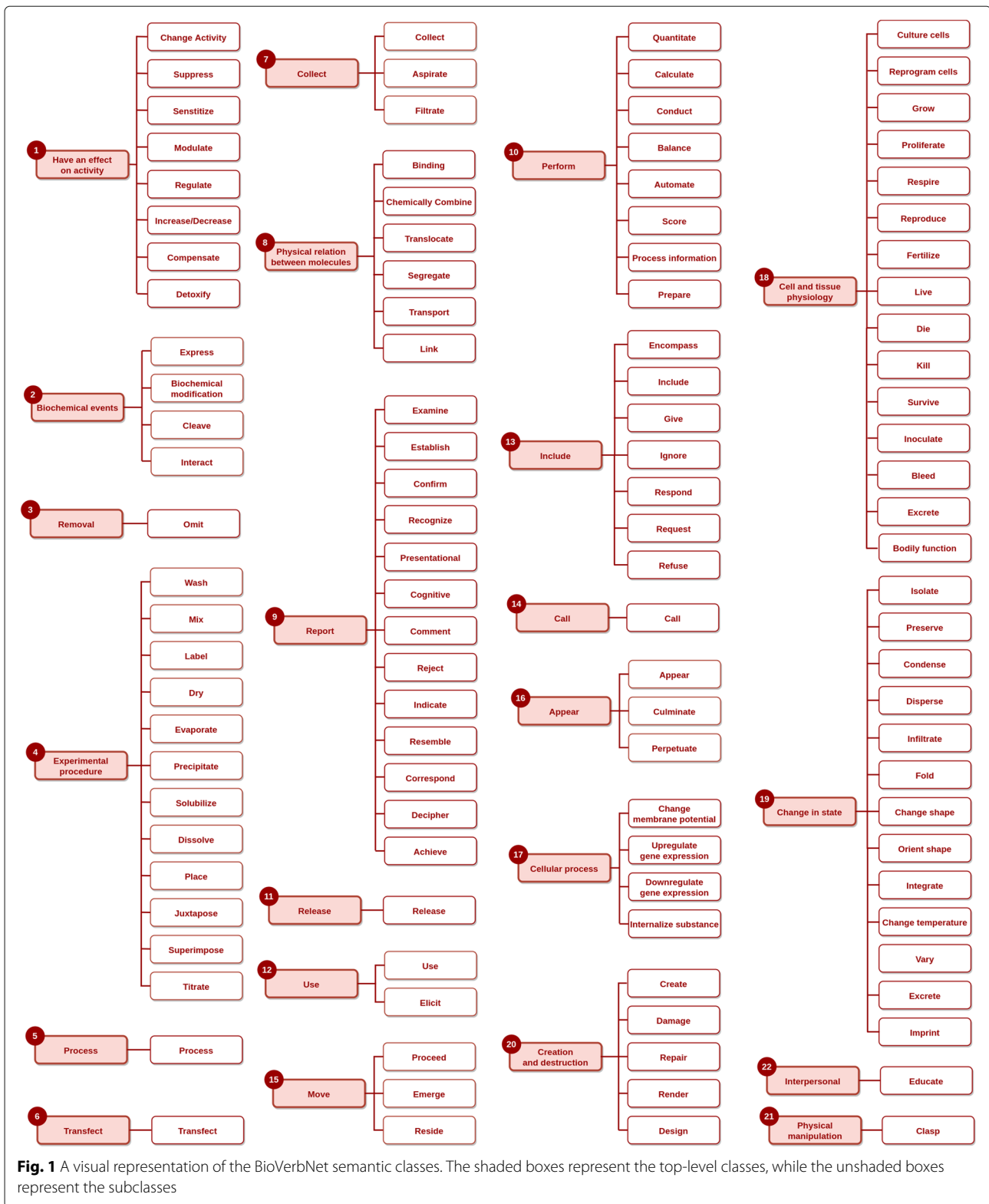


Fig. 1 A visual representation of the BioVerbNet semantic classes. The shaded boxes represent the top-level classes, while the unshaded boxes represent the subclasses

The classification approach of Chiu et al. involves the following steps. First, they use the method of Vulić et al. [48] to identify the optimal contexts for learning verb

representations in the biomedical domain, which creates a context configuration space based on dependency relations between words and subsequently applies an adapted

beam search algorithm to find the verb-specific contexts used to generate class-specific word representations. The corpus used in this step includes the PubMed Central Open Access subset [49] and the entire set of PubMed abstracts and consists of approximately 10 billion tokens and 72 million word types. The optimised representations are evaluated on a gold standard biomedical verb similarity dataset, BioSimVerb [6], and are shown to significantly outperform the baseline model (a skip-gram model with negative sampling (SGNS) without class-specific contexts).

Next, the learned representations optimised for biomedical verbs are employed as features for verb classification. To this end, Chiu et al. [13] use a small manually curated VerbNet-style classification of 192 biomedical verbs [50] and expand it with 957 new candidate verbs, assigned to the existing classes using the Nearest Centroid Classifier. The new candidates are derived from BioSimVerb based on their frequency in biomedical journals covering 120 subdomains of biomedicine, which guarantees wide coverage of the resultant classification. The final resource comprises 1149 verbs assigned to 50 classes and additionally provides the most frequent dependency contexts (and example sentences) for each verb.

Manual verification and extension

Manual verification of the automatic classes from Chiu et al. [13] was carried out in collaboration by two experts, a biologist and a linguist, with postgraduate level of training in their respective fields¹. They were provided with the original class names and member verbs from the gold standard of Korhonen et al. 2006, the new verbs assigned to the original classes by the classifier, and the set of 10 most frequent dependency contexts for each verb in the sample. For each of the dependency contexts, three example sentences extracted from the corpora and demonstrating the usage of the target verb were also provided.

The goal of the verification process was to examine the automatically generated candidates and check whether they satisfied class membership criteria. Since the aim of this work was to produce a VerbNet-style classification, we adopted an analogous definition of a class as a grouping of verbs with shared semantics and syntactic behaviour, based on the assumption that a verb's syntactic properties, such as the types of arguments it selects, inform its semantics. This rationale constitutes the foundation of Levin's [20] (1993) classification of English verbs, which has been extended and refined to create VerbNet.

¹Due to the complementary nature of their expertise, the two experts carried out the verification collaboratively and all decisions were made jointly.

The verification procedure involved the following steps. First, for each class, the automatically generated new candidate verbs were examined with respect to the original members to ensure semantic coherence; only verbs with meanings similar to the original members were kept. Then, the new candidates were reviewed in terms of their syntactic behaviour, exemplified by the dependency contexts extracted from biomedical corpora, and the class was further refined to include the subset of verbs characterised by common syntactic patterns. Based on the examination of semantic and syntactic properties of the new candidates with regard to the original members, for each new candidate the experts decided if the verb was correctly assigned; otherwise, it was (a) reassigned to another existing class, or (b) a new class was created and the verb in question was assigned to it. During manual verification, some of the original classes were split into smaller subclasses, defined by shared semantics and syntactic behaviour of their members. However, we preserved the information about overlapping semantic properties of groups of subclasses by adopting a two-tier classification structure, with general top-level classes encompassing several narrower subclasses with related semantics. At the end of this process, any verbs which could not be correctly assigned to any of the classes were discarded as noise. Table 1 reports the relevant data statistics, including the number of original classes, the number of reassignments, the number of new classes created in the verification stage, and the number of noisy candidates which were ultimately excluded from the classification.

Comparison with VerbNet

One of the important motivations behind the creation of BioVerbNet is the discrepancy between the general and biomedical domains in terms of language use and words' distributional properties. This concerns domain-specific words which are frequent in biomedical texts and absent or very rare in the general domain (e.g., *deacetylate*, *hydroxylate*), as well as words which are common in both domains, but are used in a very narrow, domain-specific sense in the biomedical domain (e.g., *prune*, *perturb*). In order to examine these discrepancies further, we comparatively analysed the newly created BioVerbNet classes and the existing VerbNet, focusing on the verbs appearing in both.

The results of this analysis support our assumptions: most verbs present in BioVerbNet are missing in VerbNet, and only 39 out of 117 BioVerbNet sub-classes contain one or more VerbNet verbs. In total, 63 of 693 class-assigned BioVerbNet verbs are also present in VerbNet. As expected, sub-classes containing highly biomedical-specific verbs have little or no overlap with VerbNet. For example, sub-class 2.2.1 'Biochemical modification' contains 20 verbs, all of which are specific to BioVerbNet.

Table 1 Assignment of verbs into classes and sub-classes

	Top-level classes	Sub-classes	Verbs
Automatically assigned verbs	16	48	283 (29.4%)
Manually assigned verbs	-	-	410 (42.7%)
Re-assigned within original sub-classes	-	-	104
Assigned to new sub-classes created within original top-level classes	-	30	93
Assigned to new top-level classes	6	39	213
Total assigned verbs	22	117	693 (72.1%)
Non-assigned verbs	-	-	268 (27.9%)
Total verbs	-	-	961 (100%)

Other BioVerbNet sub-classes contain a proportion of verbs which are also present in VerbNet, for example, 9.2.2 ‘Cognitive’ contains 39 verbs of which 5 are shared with VerbNet. In all sub-classes, shared verbs form no more than a minority. Moreover, some verbs present in BioVerbNet are characterised by senses that differ from typical usage in the general domain. In Table 2 we provide selected examples.

Semantic and syntactic annotation

In the original VerbNet classification, each class is accompanied by a set of syntactic descriptions (i.e., syntactic frames), which illustrate the possible surface realisations of a verb’s arguments. For each type of syntactic frame, a sentence example of usage is provided, along with the semantic roles of the verb’s arguments and corresponding semantic predicates (e.g., ‘motion’, ‘has_state’) with temporal and causal subevent structure. In this work, we focus annotation efforts on two main components: semantic roles and syntactic frames. For each BioVerbNet class generated in the manual verification stage, we identify the subset of shared syntactic contexts licensed by all members. Next, we annotate the class-specific syntactic structures with semantic roles and syntactic constituents. We describe the methodology adopted in each step in the following sections.

Context selection

To identify the subset of shared argument structures and syntactic contexts for each class we utilise the dependency contexts extracted from the PubMed corpus by Chiu et al. [13], which were used in the Manual Verification stage as class membership criteria. Since the dependency contexts of Chiu et al. [13] were automatically generated, they sometimes contained noise and parsing errors. Additionally, some of the contexts were redundant or uninformative with regard to the verb’s argument structure. For example, context types *obj* and *subj#obj* provided redundant information about the verb’s transitive behaviour. Moreover, some dependency contexts included conjunctions (e.g., *and*) or adjunct adverbial and prepositional phrases (e.g., *Physiologic mechanisms regulate hemodynamics during exercise and in heart failure*), which are optional elements and therefore are not considered as characteristic of a given verb’s behaviour. We adopted an iterative context identification protocol, in which for each class, for each class member, we examined the set of 10 most frequent dependency contexts and filtered those redundant or uninformative; then, the remaining contexts were checked against the class members by substituting them one by one into a given context. Only the contexts in which all the class members could participate were kept. After the set of 10 most frequent dependency contexts

Table 2 Examples of common verbs with senses specific to the biological sciences

	Sense	Example
<i>silence</i>	To inactivate expression of a gene.	Eukaryotic cells express small noncoding RNAs to silence target genes
<i>dampen</i>	To suppress the immune response.	Propathogenic cells dampen the early T cell response
<i>scavenge</i>	To combine with and remove reactive oxygen species.	Antioxidant properties of plants scavenge reactive oxygen species
<i>prime</i>	To present antigen to naïve lymphocytes, causing them to differentiate.	These antigens may prime an immune response
<i>reprogram</i>	To transform one cell type into a different cell type.	Mash1 and Brn2 reprogram fibroblasts into neurons
<i>imprint</i>	To inactivate expression of a gene through methylation.	A period of stimulation could imprint on a T cell a “biochemical memory”
<i>divide</i>	To undergo cell division into two or more daughter cells.	Cultures of <i>Tetrahymena pyriformis</i> were induced to divide synchronously
<i>isolate</i>	To extract a cell population or substance in a pure form.	We used soft agar to isolate phototrophic bacteria

was reviewed for the first verb belonging to a given class, the procedure was repeated for all the remaining class members.

Semantic role labeling

The first step of the manual annotation process involved annotating the class-specific syntactic contexts identified in the previous step with semantic roles. Also known as thematic or theta roles, semantic roles describe the underlying relationship between a participant of an event and the main verb in a clause. They capture the differences in verb meaning as reflected in the expression of its arguments and thus provide important generalisations about the interplay of verbs' semantic and syntactic behaviour, which contribute to the semantic-syntactic mapping.

While consensus has not been reached on the semantic role inventory [51–53], most approaches agree on a number of principal roles and their corresponding definitions, such as *Agent*, the instigator of the action denoted by the predicate, or *Patient*, the entity undergoing the effect of the event [7, 18, 54, 55]. To ensure alignment with VerbNet, we adopted the same set of roles and definitions. However, the discrepancies between the verbs' common usages in the general and the biomedical domains posed a number of challenges, which required a careful revision of the role assignment criteria, in view of the characteristic properties of biomedical verbs.

Challenges and domain-specific roles

The first important difference between the two domains, and consequently the two lexicons, lies in the nature of typical arguments. In VerbNet, the annotated examples predominantly feature canonical role-argument pairings, e.g., animate, intentional Agents, inanimate, concrete objects as Instruments, or human Experiencers. In the biomedical domain, the typical event participants are biological and chemical entities, such as cells, chemical reactions, or hormones. In the sentence *NKT cells mediate autoreactivity*, cells are not only animate, but they also interact with each other and their environment. However, they are not intentional, which is one of the criteria of

agency in VerbNet: *Actor in an event who initiates and carries out the event intentionally or consciously, and who exists independently of the event*. Similar considerations involve organs, tumours, or bacteria, which commonly take on the agentive role in biomedical texts. In light of the widespread nature of this phenomenon, we propose a new role, Bio-Agent. We posit it as a subtype of Causer, i.e., *an Actor in an event that initiates and effects the event and that exists independently of the event*, constrained to being a biological process, event or entity as a selectional restriction. A Bio-Agent may deploy a biochemical messenger as an inanimate Instrument, as in the sentence *Endothelial cells release substances that hyperpolarize vascular smooth muscle*, where 'Endothelial cells' are Bio-Agents and 'substances' are Instruments. In the sentence *Poxviruses deploy genomic accordions to adapt rapidly*, 'Poxviruses' are Bio-Agents and 'genomic accordions' constitute a biological mechanism functioning as an Instrument. Given the co-presence of canonical Agents (e.g., human actors) in biomedical texts, adopting the role of Bio-Agent helps capture important differences in the characteristic properties of these two types of arguments. Consequently, it allows discrimination between verbs which only permit one type of Agent, thus enabling a more fine-grained classification.

A second difference is that some verbs present in both VerbNet and BioVerbNet are characterised by differences in their typical usage and corresponding semantic roles (Table 3).

In VerbNet, *settle* either describes cognitive agreement and is classified with verbs such as *communicate*, *concur* and *compromise* in the Settle class, or belongs to the Lodge class in the sense 'to go and live somewhere' with members such as *dwell*, *reside* and *stay*. In the sentence *The couple settled there*, 'the couple' takes on the role of Agent as the instigator of the action. In biomedical texts, the first (cognitive) sense occurs in analogous contexts, while the second describes the physical movement of objects towards a stationary state. However, these objects are no longer agentive. In BioVerbNet, *settle* is placed in sub-class

Table 3 Examples of differences in semantic roles of the arguments of the same verb (underlined) in VerbNet and BioVerbNet

Source	Example sentence	Verb Frame
VerbNet	The couple <u>settled</u> there	Agent V Location
BioVerbNet	Most parasites <u>settle</u> within this area	Patient V {in} Location
VerbNet	The gardener <u>grew</u> that acorn into an oak tree	Agent V Patient {into} Product
BioVerbNet	Alga-free paramecia and symbiotic algae can <u>grow</u> independently	Agent {and} Co-Agent {can} V ADV
VerbNet	He <u>responded</u> to my call	Agent V Theme
BioVerbNet	Plants <u>respond</u> to damage	Agent V Source
VerbNet	The secretary <u>transcribed</u> the speech	Agent V Theme
BioVerbNet	These viruses <u>transcribe</u> their genomes in the nuclei of infected cells	Bio-Agent V Patient {in} Location

19.2.1 ‘Condense’, containing 16 members that include *sediment*, *coalesce* and *agglutinate*. In the sentence *Most parasites settle within this area*, the phrase ‘Most parasites’ takes on the role of Patient since it experiences the effect of the event.

The usage of the verb *transcribe* also differs between VerbNet and BioVerbNet. In VerbNet, *transcribe* describes the copying of speech or text and is placed in a group of 20 members that includes the verbs *chronicle*, *photocopy* and *record*. In biomedical texts, however, *transcribe* is typically used in the sense of DNA transcription, an active process whereby the DNA double helix is unzipped and a complementary strand of mRNA synthesized. In BioVerbNet, *transcribe* is placed in sub-class 17.2.1 ‘Upregulate gene expression’, containing 6 members that include *transactivate*, *upregulate* and *derepress*. The usage of *transcribe* in VerbNet involves a human Agent and the object of the verb is unchanged, therefore a Theme, as in *The secretary transcribed the speech* (Agent V Theme). In BioVerbNet, the agentive role is taken by either a cellular Bio-Agent or a biochemical Force, and the object of the verb is materially altered, therefore a Patient, as in *These viruses transcribe their genomes in the nuclei of infected cells* (Bio-Agent V Patient {in} Location).

Syntactic frame annotation

The second step of the annotation process consisted in annotating the characteristic frames for each subclass with syntactic constituents. For each frame, we identified word groups functioning as a single unit in the syntactic structure of the sentence (e.g., *NP*, *VP*, *AdjP*). These syntactic patterns largely overlap with those used in VerbNet. However, unlike in VerbNet, we have included for certain classes passive constructions and dependent clauses containing the target verb when those syntactic patterns typify the use of those verbs in biomedical text. Table 4 provides examples of syntactic annotation selected from the complete resource.

Utility and discussion

Evaluation

The objective of this evaluation is to apply a standard retrofitting method to change the vector-space of the pre-trained word embeddings to better capture the semantics represented by the BioVerbNet classes [56]. We apply retrofitting to our pretrained embeddings (we use the embeddings pre-trained by Chiu et al. [57]). We base our retrofitting approach on the method proposed by Faruqui et al. [58]. Given any pretrained vector-space representation, the main idea of retrofitting is to pull words which

Table 4 Examples of syntactic annotation (verb class members underlined)

Verb sub-class	Example sentence	Syntactic annotation
1.1.2 Suppress	Amine groups <u>quench</u> the excited fluorophore	NP V NP
1.3.0 Increase/Decrease	Hormonal stimuli <u>decline</u>	NP V
	Antibody levels <u>decline</u> rapidly	NP V ADVP
2.2.2 Cleave	The activated caspases <u>truncate</u> procaspase-3	NP V NP
2.3.0 Interact	The conjugated salts <u>chop</u> the cell membrane into pieces	NP V NP PP
	Both drug classes <u>synergize</u>	NP V
4.1.1 Wash	Estrogen may <u>synergize</u> with nonaromatizable androgens	NP V PP
	Subepithelial mucous gland secretions <u>clean</u> the valvular crypts	NP V NP
4.2.0 Precipitate	Specific antisera <u>coprecipitate</u> IGFBP-5	NP V NP
	VITF-A and the viral capping enzyme <u>copurify</u> to near homogeneity	NP V PP
8.1.2 Chemically combine	Nonfunctional receptors could not <u>dimerize</u>	NP V
	Curcumin can <u>chelate</u> metal ions	NP V NP
	Lomefloxacin can <u>chelate</u> with heavy metals	NP V PP
9.5.0 Decipher	We <u>comprehend</u>	NP V
	We <u>decipher</u> the molecular determinants	NP V NP
10.2.0 Score	We <u>classify</u> these diseases as immunodeficiencies	NP V NP PP
	Clinicians <u>classify</u> the patient correctly	NP V NP ADVP
17.2.2 Downregulate gene expression	HDAC4 and MEF2C <u>downmodulate</u> c-jun promoter activity	NP V NP
	MicroRNAs <u>silence</u> the expression of target genes post-transcriptionally	NP V NP PP ADVP
20.1.3 Repair	Hematopoietic stem cells can <u>reconstitute</u> the bone marrow	NP V NP
	Adult zebra fish <u>regenerate</u> their caudal fin following partial amputation	NP V NP PP

Table 5 Summary statistics of the Hallmarks of Cancer (HOC) and the Chemical Exposure Assessment (CEA) datasets

	HOC		CEA	
	Document	Sentence	Document	Sentence
Train	1,303	12,279	2,555	25,307
Dev	183	1,775	384	3,770
Test	366	3,410	722	7,100
Total	1,852	17,464	3,661	36,177

are connected in relation to the provided semantic lexicon closer together in the vector space. The main objective function to minimize in the retrofitting model is expressed as

$$\sum_{i=1}^{|V|} \left(\alpha_i \|\vec{v}_i - \tilde{v}_i\| + \sum_{(i,j) \in S} \beta_{ij} \|\vec{v}_i - \vec{v}_j\| \right) \quad (1)$$

where $|V|$ represents the size of the vocabulary, \vec{v}_i and \vec{v}_j correspond to word vectors in a pretrained representation, and \tilde{v}_i represents the output word vector. S is the input lexicon represented as a set of linguistic constraints—in our case, they are pairs of word indices, denoting the pairwise relations between member verbs in each BioVerbNet class. For example, a pair (i, j) in S implies that the i th and j th words in the vocabulary V belong to the same verb class.

The values of α_i and β_{ij} are predefined and control the relative strength of associations between members. We follow the default settings for these values as stated in the authors' work by setting $\alpha = 1$ and $\beta = 0.05$ in all of the experiments. To minimize the objective function for a set of starting vectors \vec{v} and produce retrofitted vectors \tilde{v} , we run stochastic gradient descent (SGD) for 20 epochs. An implementation of this algorithm has been published online by the authors;² we used this implementation in this evaluation.

We evaluate our word representations using two established biomedical datasets for text classification: the Hallmarks of Cancer (HOC) [59, 60] and the Chemical Exposure Assessment (CEA) taxonomy [16]. We evaluate each based on their document-level (Pubmed abstract) and sentence-level classifications, where zero or more predefined labels can be assigned for both of these tasks.

The Hallmarks of Cancer depicts a set of interrelated biological factors and behaviours that enable cancer to thrive in the body. Introduced by Weinberg and Hanahan [15], it has been widely used in biomedical NLP, including as part of the BioNLP Shared Task 2013, "Cancer Genetics task" [61]. Baker et al. [59, 60] have released an expert-annotated dataset of cancer hallmark classifications for both sentences and documents in PubMed. The data consists of multi-labelled documents and sentences using a taxonomy of 37 classes.

The Chemical Exposure Assessment taxonomy, introduced by Larsson et al. [16], is an annotated dataset for the classification of text (documents or sentences) concerning chemical risk assessments. The taxonomy of 32 classes is divided into two branches: one relates to assessment of exposure routes (ingestion, inhalation, dermal absorption, etc.) and the second to the measurement of exposure biomarkers (biomonitoring). Table 5 details basic statistics for each dataset.

We input the retrofitted vectors into a baseline neural network model; we use the convolutional neural network (CNN) model proposed by Kim [62] for text classification tasks. An implementation of this model that was used on both the Hallmarks of Cancer task and the Chemical Exposure Assessment task has been published by Baker et al. [63]; we use this implementation in our experiment. The input to the model is an initial word embedding layer that maps input texts into matrices, which is then followed by convolutions of different filter sizes, 1-max pooling, and finally a fully-connected layer leading to an output Softmax layer predicting labels for text. Model hyperparameters and the training setup are summarized in Table 6.

For both tasks, we use the embeddings³ by Chiu et al. [57] without retrofitting as a control baseline, and we evaluate two variations of the BioVerbNet verb classes, the 22 top-level classes, and the 117 subclasses.

Performance is evaluated using the standard precision, recall, and F_1 -score metrics of the labels in the model using the one-vs-rest setup: we train and evaluate K independent binary CNN classifiers (*i.e.* a single classifier per

Table 6 Hyper-parameters used in our convolutional neural network

Parameters	Values
Vector dimension	200
Filter sizes	3,4 and 5
Number of filters	300
Dropout probability	0.5
Minibatch size	50
Input size (in tokens)	500 (documents), 100 (sentences)

²<https://github.com/mfaruqui/retrofitting>³<https://github.com/cambridgeltl/BioNLP-2016>

Table 7 Evaluation results for the Hallmarks of Cancer task (HOC) text classification task

Model	Document classification			Sentence classification		
	Precision	Recall	F_1	Precision	Recall	F_1
Baseline (no retrofitting)	77.8	51.7	62.1	56.8	30.7	39.9
22-classes retrofitted	74.4	62.1	67.7*	49.1	35.8	41.4*
117-subclasses retrofitted	74.8	62.5	68.1*	48.6	35.2	40.8*

The Baseline model is a skip-gram model without any retrofitting. All figures are micro-averages expressed as percentages (Bold denotes the best F_1 -score, * denotes statistically significant scores with respect to the baseline)

class with the instances of that class as positive samples and all other instances as negatives). Due to their random initialization, we repeat each CNN experiment 20 times and report the mean of the evaluation results to account for variances in neural networks. To address overfitting in the CNN, we use early stopping; testing only the model that achieved the highest results on the development dataset. We apply a two-tailed t-test with $\alpha = 0.05$ on the averaged output in comparison with the baseline model.

The results of our valuations are summarised in Table 7 for the HOC task, and Table 8 for the CEA task. We can observe that in both classification tasks, and at both levels of text classification (document and sentence), the retrofitted models outperformed the baseline models with significant results. The more fine-grained 117 subclasses retrofitting improved the document-level classification for both tasks more than the top-level verb classes, whereas for sentence classification the opposite is observed. For the HOC task, Recall benefited substantially from the retrofitting process, whereas for the CEA task both Precision and Recall improved slightly compared to the baseline. The reason behind the difference is likely because the HOC dataset contains classes that are very sparse (with only a small number of examples), and therefore recall would increase more substantially for these classes at the cost of precision; this has also been observed in prior work with the HOC task [56, 63, 64].

These results demonstrate the utility of BioVerbNet for specialising distributional word embeddings to better capture the properties of verbs in biomedicine and reveal its potential to aid NLP models in tackling domain-specific tasks where accurate verb processing is important.

Conclusions

This paper introduces BioVerbNet, the first semantic-syntactic classification of biomedical verbs. The resource groups verbs occurring in the PubMed corpus based on shared meaning and syntactic behaviour into 22 top-level and 117 fine-grained classes, each described by a set of characteristic syntactic frames, annotated with semantic roles. To construct BioVerbNet, we started from the output of a neural classification method specialised for biomedical verbs [13], which subsequently underwent manual revision and refinement, as well as semantic-syntactic annotation, by domain and linguistics experts. The resource provides VerbNet-style information for members of each class, including the characteristic syntactic contexts in which they appear and the typical semantic roles taken by their arguments.

BioVerbNet fills the gap in computational lexical resources targeting biomedical verbs currently available and promises to support future work in biomedical NLP. Our evaluation experiments on the task of text classification demonstrated that BioVerbNet can be readily used to support natural language processing models in biomedical tasks. We showed that class membership information from BioVerbNet can be successfully leveraged by retrofitting pretrained word embeddings so that verbs sharing the same BioVerbnet class, and therefore semantic and syntactic behaviour, are pulled closer together in the embedding space. Our retrofitted embeddings outperformed the baseline models by a significant margin on two datasets, Hallmarks of Cancer and Chemical Exposure Assessment taxonomy. Moreover, the resource provides detailed, manually-curated semantic-syntactic annotations for each class, which offer insights into the

Table 8 Evaluation results for the Chemical Exposure Assessment (CEA) text classification task

Model	Document classification			Sentence classification		
	Precision	Recall	F_1	Precision	Recall	F_1
Baseline (no retrofitting)	89.5	87.1	88.3	66.2	62.8	64.5
22-classes retrofitted	89.9	87.5	88.7*	67.3	62.1	64.6
117-subclasses retrofitted	89.2	88.6	88.9*	66.3	60.3	63.2*

Baseline model is a skip-gram model without any retrofitting. All figures are micro-averages expressed as percentages (Bold denotes the best F_1 -score, * denotes statistically significant scores with respect to the baseline)

domain-specific properties of biomedical verbs and can support researchers in developing models capable of nuanced handling of the syntactic and semantic properties of verbs in biomedical texts.

Future work

BioVerbNet includes 961 biomedical verbs sampled from PubMed corpora, making it the largest lexicon of this kind available in biomedicine. In future work, it can be further extended to cover less frequent verbs and additional classes. Moreover, given that the annotation style in BioVerbNet follows that used in VerbNet, the two resources can be linked at the level of individual verbs appearing in both, thus providing richer information for each entry and enabling easier comparisons of verb behaviour in both domains.

In future work, we will further explore the potential of BioVerbNet to support state-of-the-art NLP systems in solving biomedical tasks. Given the success of BioBERT, we will use our new resource to probe its ability to capture verbal meaning in biomedical texts and compare its performance against our best performing embeddings retrofitted to BioVerbNet class membership information. Moreover, we will investigate the potential of injecting knowledge about biomedical verbs from BioVerbNet into large pretrained encoders to further boost their verbal reasoning capacity in biomedicine. To support future endeavours in biomedical NLP we make our resource freely available to the community at <https://github.com/cambridgeitl/bioverbnet>.

Abbreviations

NLP: Natural Language Processing; UMLS: Unified Medical Language System; BERT: Bidirectional Encoder Representations from Transformers; BOW: Bag-Of-Words; SGNS: Skip-gram model with negative sampling; LSTM: Long short-term memory; ELMo: Embeddings from Language Models

Acknowledgements

Not applicable

Authors' contributions

OM: Designed the resource construction protocol and annotation guidelines, carried out manual verification of the classification and semantic-syntactic annotation. CC: Performed manual verification and extension of the automatic classes, refined the two levels of classification, carried out semantic-syntactic annotation. SB: Performed the experiments demonstrating the utility of the new resource in boosting model performance in document- and sentence-level classification in biomedicine. JB: Provided guidance on the data, domain and experiments. SWB: Verified the classification and provided guidance on the semantic and syntactic annotation. AK and MP: Supervised the work and provided guidance on the construction of the resource. All authors read and approved the final manuscript.

Funding

This work is supported by the ERC Consolidator Grant LEXICAL [grant number 648909].

Availability of data and materials

The dataset created in this study is available on Github, <https://github.com/cambridgeitl/bioverbnet>. The datasets used for evaluation are publicly available at: Exposure Assessment taxonomy <http://dx.doi.org/10.6084/m9.figshare.4668229> and Hallmarks of Cancer Corpus <https://github.com/sb895/Hallmarks-of-Cancer>.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

¹Language Technology Laboratory, MMLL, University of Cambridge, 9 West Road, CB39DB Cambridge, UK. ²Department of Future Technologies, University of Turku, Vesilinnantie 5, 20500 Turku, Finland. ³Department of Linguistics, University of Colorado Boulder, 295 UCB, 80309-0295 Boulder, Colorado, USA.

Received: 28 January 2021 Accepted: 1 July 2021

Published online: 15 July 2021

References

- Schuyler PL, Hole WT, Tuttle MS, Sherertz DD. The UMLS Metathesaurus: representing different views of biomedical concepts. *Bull Med Libr Assoc*. 1993;81(2):217.
- Ananiadou S, McNaught J. *Text Mining for Biology and Biomedicine*. London: Artech House; 2006.
- Venturi G, Montemagni S, Marchi S, Sasaki Y, Thompson P, McNaught J, Ananiadou S. Bootstrapping a verb lexicon for biomedical information extraction. In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer; 2009. p. 137–48. https://doi.org/10.1007/978-3-642-00382-0_11.
- Tan H. A system for building FrameNet-like corpus for the biomedical domain. In: *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*. Association for Computational Linguistics; 2014. p. 46–53. <https://doi.org/10.3115/v1/w14-1107>.
- Mondal A, Das D, Cambria E, Bandyopadhyay S. WME 3.0: An enhanced and validated lexicon of medical concepts. In: *Proceedings of the 9th Global WordNet Conference (GWC)*. Nanyang Technological University (NTU): Global Wordnet Association; 2018. p. 10–6. <https://aclanthology.org/2018.gwc-1.2>.
- Chiu B, Pyysalo S, Vulić I, Korhonen A. Bio-SimVerb and Bio-SimLex: Wide-coverage evaluation sets of word similarity in biomedicine. *BMC Bioinformatics*. 2018;19(1):33.
- Kipper K, Korhonen A, Ryant N, Palmer M. A large-scale classification of English verbs. *Lang Resour Eval*. 2008;42(1):21–40.
- Brown SW, Dligach D, Palmer M. VerbNet class assignment as a WSD task. In: *Proceedings of the Ninth International Conference on Computational Semantics*. Association for Computational Linguistics; 2011. p. 85–94. <https://aclanthology.org/W11-0110>.
- Giuglea A-M, Moschitti A. Semantic role labeling via FrameNet, VerbNet and PropBank. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. Sydney: Association for Computational Linguistics; 2006. p. 929–36. <https://doi.org/10.3115/1220175.1220292>.
- Schmitz M, Bart R, Soderland S, Etzioni O, et al. Open language learning for information extraction. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island: Association for Computational Linguistics; 2012. p. 523–34. <https://aclanthology.org/D12-1048>.
- Lippincott T, Rimell L, Verspoor K, Korhonen A. Approaches to verb subcategorization for biomedicine. *J Biomed Inform*. 2013;46(2):212–27.
- Rimell L, Lippincott T, Verspoor K, Johnson HL, Korhonen A. Acquisition and evaluation of verb subcategorization resources for biomedicine. *J Biomed Inform*. 2013;46(2):228–37.
- Chiu B, Majewska O, Pyysalo S, Wey L, Stenius U, Korhonen A, Palmer M. A neural classification method for supporting the creation of BioVerbNet. *J Biomed Semant*. 2019;10(1):2.
- The Pubmed Central Open Access Subset. 2017. <http://www.pubmedcentral.nih.gov/about/openftlist.html>. Accessed 5 Sept 2017.
- Weinberg R, Hanahan D. The hallmarks of cancer. *Cell*. 2000;100(1):57–70.

16. Larsson K, Baker S, Silins I, Guo Y, Stenius U, Korhonen A, Berglund M. Text mining for improved exposure assessment. *PLoS ONE*. 2017;12(3): 0173132. <https://doi.org/10.6084/m9.figshare.4668229>.
17. Fellbaum C, (ed). *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press; 1998.
18. Baker CF, Fillmore CJ, Lowe JB. The Berkeley FrameNet project. In: *Proceedings of COLING*; 1998. <http://aclweb.org/anthology/C98-1013>.
19. Kingsbury PR, Palmer M. From TreeBank to PropBank. In: *LREC*. Luxembourg: European Language Resources Association (ELRA); 2002. p. 1989–93.
20. Levin B. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press; 1993.
21. Rios M, Aziz W, Specia L. TINE: A metric to assess MT adequacy. In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh: Association for Computational Linguistics; 2011. p. 116–122. <https://aclanthology.org/W11-2112>.
22. Shi L, Mihalcea R. Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In: *Computational linguistics and intelligent text processing*. Berlin: Springer; 2005. p. 100–111.
23. Dang HT. *Investigations into the role of lexical semantics in word sense disambiguation*. 2004.
24. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: *Proceedings of NIPS*; 2013. p. 3111–3119. <https://dl.acm.org/doi/10.5555/2999792.2999959>.
25. Chiu B, Baker S. Word embeddings for biomedical natural language processing: A survey. *Lang Linguist Compass*. 2020;14(12):12402.
26. Phan MC, Sun A, Tay Y. Robust representation learning of biomedical names. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics; 2019. p. 3275–3285. <https://doi.org/10.18653/v1/P19-1317>.
27. Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. In: *Proc EMNLP*; 2014. p. 1532–43.
28. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. *Trans Assoc Comput Linguist*. 2017;5:135–46.
29. Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou S. Distributional semantics resources for biomedical text processing. In: *Proceedings of LBM*; 2013. p. 39–44. <http://lbm2013.biopathway.org/lbm2013proceedings.pdf>.
30. Stoeckel M, Hemati W, Mehler A. When specialization helps: Using pooled contextualized embeddings to detect chemical and biomedical entities in Spanish. In: *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*. Hong Kong: Association for Computational Linguistics; 2019. p. 11–5. <https://doi.org/10.18653/v1/D19-5702>.
31. Chen Q, Lee K, Yan S, Kim S, Wei C-H, Lu Z. BioConceptVec: Creating and evaluating literature-based biomedical concept embeddings on a large scale. *PLoS Comput Biol*. 2020;16(4):1007617.
32. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci Data*. 2019;6(1):1–9.
33. Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, Kingsbury P, Liu H. A comparison of word embeddings for the biomedical natural language processing. *J Biomed Inform*. 2018;87:12–20.
34. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans: Association for Computational Linguistics; 2018. p. 2227–37. <https://doi.org/10.18653/v1/N18-1202>.
35. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis: Association for Computational Linguistics; 2019. p. 4171–4186. <https://doi.org/10.18653/v1/N19-1423>.
36. Jin Q, Dhingra B, Cohen W, Lu X. Probing biomedical embeddings from language models. In: *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*. Minneapolis: Association for Computational Linguistics; 2019. p. 82–89. <https://doi.org/10.18653/v1/W19-2011>.
37. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234–40.
38. Lyu C, Chen B, Ren Y, Ji D. Long short-term memory RNN for biomedical named entity recognition. *BMC Bioinformatics*. 2017;18(1):462.
39. Nentidis A, Krithara A, Bougiatioti K, Paliouras G, Kakadiaris I. Results of the sixth edition of the BioASQ challenge. In: *Proceedings of the 6th BioASQ Workshop A Challenge on Large-scale Biomedical Semantic Indexing and Question Answering*. Brussels: Association for Computational Linguistics; 2018. p. 1–10. <https://doi.org/10.18653/v1/W18-5301>. <https://www.aclweb.org/anthology/W18-5301>.
40. Lim S, Lee K, Kang J. Drug-drug interaction extraction from the literature using a recursive neural network. *PLoS ONE*. 2018;13(1):0190926.
41. Zhu Y, Li L, Lu H, Zhou A, Qin X. Extracting drug-drug interactions from texts with BioBERT and multiple entity-aware attentions. *J Biomed Inform*. 2020;106:103451. <https://doi.org/10.1016/j.jbi.2020.103451>.
42. Gondane S. Neural network to identify personal health experience mention in tweets using BioBERT embeddings. In: *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*. Florence: Association for Computational Linguistics; 2019. p. 110–3. <https://doi.org/10.18653/v1/W19-3218>.
43. Das D, Katyal Y, Verma J, Dubey S, Singh A, Agarwal K, Bhaduri S, Ranjan R. Information retrieval and extraction on covid-19 clinical articles using graph community detection and bio-BERT embeddings. In: *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Association for Computational Linguistics; 2020. <https://aclanthology.org/2020.nlpcovid19-acl.7>.
44. Vlachos A, Korhonen A, Ghahramani Z. Unsupervised and constrained Dirichlet process mixture models for verb clustering. In: *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*. Association for Computational Linguistics; 2009. p. 74–82. <https://dl.acm.org/doi/10.5555/1705415.1705425>.
45. Joanis E, Stevenson S, James D. A general feature space for automatic verb classification. *Nat Lang Eng*. 2008;14(3):337–67.
46. Sun L. *Automatic induction of verb classes using clustering*. PhD thesis, University of Cambridge. 2013.
47. Barak L, Fazly A, Stevenson S. Learning verb classes in an incremental model. In: *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics*; 2014. p. 37–45.
48. Vulić I, Schwartz R, Rappoport A, Reichart R, Korhonen A. Automatic selection of context configurations for improved class-specific word representations. In: *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Vancouver: Association for Computational Linguistics; 2017. p. 112–22. <https://doi.org/10.18653/v1/K17-1013>.
49. The Pubmed Central Open Access Subset. <http://www.pubmedcentral.nih.gov/about/openftlist.html>.
50. Korhonen A, Krymowski Y, Collier N. Automatic classification of verbs in biomedical texts. In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sydney: Association for Computational Linguistics; 2006. p. 345–52. <https://doi.org/10.3115/1220175.1220219>.
51. Dowty D. Thematic proto-roles and argument selection. *Language*. 1991;67(3):547–619.
52. Levin B, Hovav MR. *Argument Realization*. Cambridge: Cambridge University Press; 2005.
53. Luraghi S, Narrog H. *Perspectives on Semantic Roles*, vol. 106. Amsterdam/Philadelphia: John Benjamins Publishing Company; 2014.
54. Fillmore CJ. In: Bach E, Harms R, editors. *The case for case*. New York: Holt, Rinehart & Winston; 1968.
55. Palmer M, Gildea D, Kingsbury P. The proposition bank: An annotated corpus of semantic roles. *Comput Linguist*. 2005;31(1):71–106.
56. Chiu B, Baker S, Palmer M, Korhonen A. Enhancing biomedical word embeddings by retrofitting to verb clusters. In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence: Association for Computational Linguistics; 2019. p. 125–34. <https://doi.org/10.18653/v1/W19-5014>.
57. Chiu B, Crichton G, Korhonen A, Pyysalo S. How to train good word embeddings for biomedical NLP. In: *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. Berlin: Association for

- Computational Linguistics; 2016. p. 166–74. <https://doi.org/10.18653/v1/W16-2922>.
58. Faruqui M, Dodge J, Jauhar SK, Dyer C, Hovy E, Smith NA. Retrofitting word vectors to semantic lexicons. In: Proc. of NAACL. Denver: Association for Computational Linguistics; 2015. p. 1606?–15. <https://doi.org/10.3115/v1/N15-1184>.
59. Baker S, Silins I, Guo Y, Ali I, Högberg J, Stenius U, Korhonen A. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*. 2015;32(3):432–40.
60. Baker S, Ali I, Silins I, Pyysalo S, Guo Y, Högberg J, Stenius U, Korhonen A. Cancer Hallmarks Analytics Tool (CHAT): A text mining approach to organize and evaluate scientific literature on cancer. *Bioinformatics*. 2017;33(24):3973–81.
61. Pyysalo S, Ohta T, Ananiadou S. Overview of the cancer genetics (CG) task of BioNLP shared task 2013. In: Proceedings of the BioNLP Shared Task 2013 Workshop. Sofia: Association for Computational Linguistics; 2013. p. 58–66. <https://aclanthology.org/W13-2008>.
62. Kim Y. Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha: Association for Computational Linguistics; 2014. p. 1746–51. <https://doi.org/10.3115/v1/D14-1181>.
63. Baker S, Korhonen A. Initializing neural networks for hierarchical multi-label text classification. In: BioNLP 2017. Vancouver: Association for Computational Linguistics; 2017. p. 307–15. <https://doi.org/10.18653/v1/W17-2339>.
64. Baker S, Korhonen A, Pyysalo S. Cancer hallmark text classification using convolutional neural networks. In: BioTxtM 2016. Osaka: The COLING 2016 Organizing Committee; 2016. p. 1–9. <https://aclanthology.org/W16-5101>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

