Enhancing biomedical word embeddings by retrofitting to verb clusters

Billy Chiu[1] **Simon Baker**[1] **Martha Palmer**[2] **Anna Korhonen**[1]

[1] Language Technology Lab, University of Cambridge, 9 West Road, Cambridge, CB3 9DB, UK
[2] Department of Linguistics, University of Colorado at Boulder, Colorado, 80309–0295, USA
{hwc25,sb895,alk23}@cam.ac.uk
mpalmer@colorado.edu

Abstract

Verbs play a fundamental role in many biomedical tasks and applications such as relation and event extraction. We hypothesize that performance on many downstream tasks can be improved by aligning the input pretrained embeddings according to semantic verb classes. In this work, we show that by using semantic clusters for verbs, a large lexicon of verb classes derived from biomedical literature, we are able to improve the performance of common pretrained embeddings in downstream tasks by retrofitting them to verb classes. We present a simple and computationally efficient approach using a widely-available "off-the-shelf" retrofitting algorithm to align pretrained embeddings according to semantic verb clusters. We achieve state-of-the-art results on text classification and relation extraction tasks.

1 Introduction

Core tasks in biomedical natural language processing (BioNLP) such as relation and event extraction, text classification, syntactic and semantic parsing, natural language inference, and entailment can all benefit from rich computational lexicons containing information about the behaviour and meaning of words in biomedical texts. Verbs are especially important in many of these tasks (Cohen et al., 2008); for example, describing protein–protein interactions in biomedical text can often rely on a wide range of verbs, such as "bind," "activate," "carry," "facilitate," "interact," *etc*. in order to determine the specific type of interaction.

Lexical semantic classes for verbs can be used to abstract away from individual words, or to build a lexical structure (taxonomy) which predicts much of the behaviour of a new word by associating it with an appropriate class (Levin, 1993; Kipper et al., 2008). For example, the verbs "assess," "evaluate," "estimate," "explore," and "analyze" belong to the class *examine*, while the verbs "utilize," "employ," and "exploit" belong to the class *use*. In addition to simple synonyms of verbs, semantic classes capture similarity in their use and behaviour in text by analysing their contexts (Levin, 1993).

In the past, lexical verb classes have been successfully shown to improve the performance classifiers in a variety of tasks and down stream applications in the biomedical domain; such as relation extraction (Sharma et al., 2010), biomedical fact extraction (Rupp et al., 2010), text classification for cancer (Baker et al., 2015), biomedical discourse analysis (Cox et al., 2017), and biomedical information retrieval (Mahalakshmi, 2015).

Lexical classes are useful for their ability to capture generalizations about a range of linguistic properties (Kipper et al., 2000); our hypothesis is therefore that by retrofitting embedded word representations to semantic verb classes, semantically-similar verbs (*i.e*. member verbs within the same lexical class) like "suppress" and "inhibit" will be pulled together in vector space, whereas verbs like "collect" and "examine" will not. Consequently, this allows NLP systems to generalize away from individual verbs, alleviating the data sparseness problem of representing each verb in the corpus individually.

Retrofitting is a graph-based learning technique for using lexical relational resources to obtain higher quality semantic vectors (Faruqui et al., 2015). It is applied as a post-processing step by running belief propagation on a graph constructed from lexicon-derived relational information to update word vectors. It can be applied to any pretrained word embedding vectors. The intuition behind retrofitting is to encourage the retrofitted vectors to be similar to the vectors of related word

Proceedings of the BioNLP 2019 workshop, pages 125–134
Florence, Italy, August 1, 2019.

types and similar to their original distributional representations.

Using a standard "off-the-shelf" retrofitting algorithm, we apply the idea of retrofitting to verb clusters to two sets of widely-used pretrained embedding vectors in BioNLP (those by Pyysalo et al. (2013a) and by Chiu et al. (2016)) to obtain improved embeddings. We show that by doing nothing more than using this simple approach, we achieve state-of-the-art results on two text classification tasks (both tasks evaluated on document and sentence level classification), and a relation extraction task. We make our retrofitted embeddings freely available to the BioNLP community along with our code.[1]

The main contribution of this work is to be the first of its kind to apply verb-based retrofitting in the biomedical domain. Retrofitting has thus far only been applied for aligning vectors to Medical Subject Headings (MeSH) (Yu et al., 2016), and been validated only in an extrinsic setting. We show that with very little effort, we can achieve state-of-the art results on various downstream tasks in a range of biomedical subdomains.

This paper will first describe relevant work on retrofitting to lexical resources in BioNLP; we then briefly give an overview of two verb cluster and lexicons that we use in our methodology, and then our task-based evaluation. We end with a discussion of the evaluation results.

2 Related work

Lexical resources can be used to enrich representation models by providing them other sources of linguistics information beyond the distributional statistics obtained from corpora. In recent literature, various methods to leverage knowledge available in human- and automatically-constructed lexical resources have been proposed.

One such method involves modifying the objectives in the original representation learning procedures so that they can jointly learn both distributional and lexical information—for example, Yu and Dredze (2014) modify the CBOW objective function by introducing semantic constraints as obtained from the paraphrase database (Ganitkevitch et al., 2013) to train word representations which focus on word similarity over word relatedness.

Another class of methods incorporates lexical information into the vector representations as a post-processing procedure. The method fine-tunes the pretrained word vectors to satisfy linguistic constraints from the external resources. The method can be applied to any off-the-shelf models without requiring large corpora for (re-)training as the joint-learning models do. Among these methods, *retrofitting* (Faruqui et al., 2015) is widely used.

Given any (pretrained) vector-space representations, the goal of retrofitting is to bring closer words which are connected *via* a relation (*e.g.* synonyms) in a given semantic network or lexical resource (*i.e.* linguistic constraints). For example, Yu et al. (2016, 2017) retrofit word vector spaces of MeSH terms by using additional linkage information from the UMNSRS hierarchy to improve the representations of biomedical concepts. Building on retrofitting, Lengerich et al. (2018) generalize retrofitting methods by explicitly modelling individual linguistic constraints that are commonly found in health and clinical-related lexicons (*e.g.* causal-relations between diseases and drugs).

In theory, the joint-learning models could be as effective (or better) as those produced by fine-tuning distributional vectors. However, the performance of joint-learning models has not surpassed that of fine-tuning methods.[2] Furthermore, the joint-learning objectives are usually model-specific and are tailored to a particular model, making it difficult to use them with other methods. In this work, we will use retrofitting to incorporate our lexical features into the word representations.

3 Verb clusters

In this work, we investigate retrofitting popular word embeddings to two publicly available[3] lexicons for verb clusters. The first is composed of 192 relatively frequent verbs from a corpus of 2230 biomedical journal articles which have been hierarchically classified into three levels: 16, 34, and 50 verb classes. The three levels reflect different granularity in the semantics of the verb classes as illustrated in Figure 1. These clusters were annotated by 4 domain experts and 2 linguists, were used to create the gold standard (Korhonen

[1] Our retrofitted embeddings and code are released under an open license and can be found here: https://github.com/cambridgeltl/retrofitted-bio-embeddings

[2] The SimLex-999 home page (www.cl.cam.ac.uk/~fh295/simlex.html) lists state-of-the-art performance models, none of which have learned representations jointly

[3] https://github.com/cambridgeltl/bio-verbnet

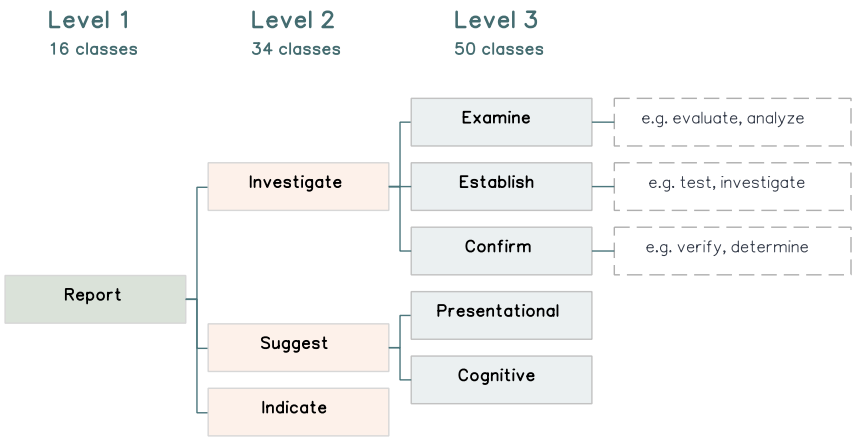

Figure 1: Examples of the verb classes introduced by Korhonen et al. (2006).

et al., 2006). We will refer to this lexicon for the remainder of this paper as the *annotated clusters*.

Chiu et al. (2019) developed a methodology to further extended the annotated clusters automatically using text from PubMed abstracts and full articles with the goal of facilitating the future creation of a BioVerbNet resource, a specialized resource similar to VerbNet (Schuler, 2005). We will refer to this lexicon for the remainder of this paper as the *expanded clusters*.

Chiu et al. (2019) use a two-step method. In the first step, the best contexts for learning biomedical verb representations are identified using a model based on skip-gram with negative sampling (SGNS). It involves first creating a context configuration space based on dependency relations between words, followed by applying an adapted beam search algorithm to search this space for the class-specific contexts, and finally using these contexts to create class-specific representations.

In the second step, the optimized representation is used to provide word features for building a verb classification. This is obtained by expanding the verbs in the annotated clusters, where the candidate verbs are selected from BioSimVerb (Chiu et al., 2018) based on their frequent occurrence in biomedical journals across 120 subdomains of biomedicine. A Nearest Centroid classifier is then used to connect the new candidates to an appropriate class. The resulting classification provides 1149 verbs assigned to the 50 classes in the original annotated clusters. For each verb, the expanded clusters lists the most frequent dependency contexts that reflect their syntactic behaviour along with example sentences.

For the rest of the work, we will investigate the use of both the annotated and expanded clusters

4 Methodology

We apply retrofitting to our default pretrained embeddings[4] The goal is to change the vector-space of the pretrained word embeddings to better capture the semantics represented by the verb classes in both the annotated and expanded clusters. These verb classes provide different levels of generalization to support various tasks, from the coarse-grained level of 16 classes to a fine-grained one of 50 classes.

We base our retrofitting method on that proposed by Faruqui et al. (2015). Given any pretrained vector-space representation, the main idea of retrofitting is to pull words which 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

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

where $|V|$ represents the size of the vocabulary, $\vec{v}_i$ and $\vec{v}_j$ corresponds to word vectors in a pretrained representation, and $\vec{\hat{v}}_i$ 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 class. For example,

[4]For our default embeddings, we use the embeddings by Chiu et al. (2016) for our text classification tasks and Pyysalo et al. (2013a) for relation extraction.

	Number of verb pairs	
	Annotated clusters	Expanded clusters
16-classes	1,774	96,998
34-classes	638	54,063
50-classes	376	50,104

Table 1: Linguistic constraint counts under each class as obtained from the Korhonen's resource and our automatically-created lexicon.

a pair (i, j) in S implies that the ith and jth 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 $\vec{\hat{v}}$, we run stochastic gradient descent (SGD) for 20 epochs. An implementation of this algorithm has been published online by the authors;[5] we used this implementation in the present work.

Table 1 shows the linguistic constraint counts under each class as derived from the two lexicons. When retrofitted against the three top levels, the member verbs at each subclass are merged with its upper class, as in the work of Faruqui et al. (2015).

5 Evaluation

We apply retrofitting to incorporate the lexical information into word representations. Then we evaluate the quality of the retrofitted-representation as features for two NLP tasks: text classification and relation classification.

5.1 Task 1: Text classification

We evaluate our word representations using two established biomedical datasets for text classification: the Hallmarks of Cancer (**HOC**) (Baker et al., 2015, 2017) and the Exposure taxonomy (**EXP**) (Larsson et al., 2017). We evaluate each based on their document-level and sentence-level classifications.

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 (2000), it has been widely used in biomedical NLP, including as part of the BioNLP Shared Task 2013, "Cancer Genetics task" (Pyysalo et al., 2013b). Baker et al. (2015, 2017) 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 Exposure taxonomy, introduced by Larsson et al. (2017), 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 bio-markers (biomonitoring). Table 2 shows basic statistics for each dataset.

	HOC		EXP	
	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

Table 2: Summary statistics of the Hallmarks of Cancer (HOC) and the Chemical Exposure Assessment (EXP) datasets.

The model follows the convolutional neural network (CNN) model proposed by Kim (2014). An implementation of this algorithm on HOC and EXP has been published by Baker and Korhonen (2017); 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 3.

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)

Table 3: Hyper-parameters used in (Baker and Korhonen, 2017).

[5] https://github.com/mfaruqui/retrofitting

For both tasks, we use the embeddings[6] by Chiu et al. (2016). 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 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 follow the authors' early stopping approach, testing only the model that achieved the highest results on the development dataset.

5.2 Task 2: Relation classification

We evaluate our retrofitted representations on the Bio-Creative VI Chemical–Protein relation extraction dataset (**CHEMPROT**) (Krallinger et al., 2017). The corpus provides mention and relation annotations for complex events related to chemical–protein interaction in molecular biology. The goal of this task is to predict whether a given chemical–protein pair is related or not, and to then verify its corresponding relation type. There are five types of relations: *Up-regulator*, *Down-regulator*, *Agonist*, *Antagonist*, and *Substrate*. The corpus is provided in the Turku Event Extraction System (TEES) XML format and are installed with the Turku Extraction System (Björne, 2014). It is parsed with the the BLLIP parser (Charniak and Johnson, 2005) with the McClosky bio-model (Mcclosky, 2010), followed by conversion of the constituency parses into dependency parses using the Stanford Tools (MacCartney et al., 2006). Table 4 summarizes key statistics for the dataset.

	Documents	Entities	Relations
Train	1,020	25,769	4,157
Dev	612	15,571	2,416
Test	800	20,829	3,458
Total	2,432	62,169	10,031

Table 4: Summary statistics of the Chemical-Protein interaction dataset (CHEMPROT).

The model follows the CNN model proposed by Björne and Salakoski (2018). We directly use their published implementation. The model input is an initial word embedding layer that maps input texts into matrices, followed by convolutions of different filter sizes and 1-max pooling, and finally a fully connected layer, leading to an output Softmax layer for predicting labels. Performance is evaluated using the standard precision, recall, and F_1-score metrics of the labels in the model. Classification is performed as multilabel classification where each example may have 0 to n positive labels. Model hyperparameters and the training setup are summarized in Table 5.

Parameters	Values
Vector dimension	200
Filter sizes	1, 3, 5 and 7
Number of filters	400 (100 of each size)
Dropout probability	0.5
Learning rate	0.001
Minibatch size	50

Table 5: Hyperparameters used by Björne and Salakoski (2018).

To account for variance in neural networks due to their random initialization, we adopt the ensemble settings used by Björne and Salakoski (2018). We train 20 models and take the n best ones ($n = 5$), ranked with their F_1-score on the development set, and use their averaged predictions. The ensemble predictions are calculated for each label as the average predicted confidence scores from all the models. We also incorporate the authors' early stopping approach where the model is trained until the development loss no longer decreases. We train for up to 500 epochs, stopping once validation loss has no longer decreased for 10 consecutive epochs. To focus on the effect of verb classes on biomedical representations, we experiment with word representations induced on biomedical texts; this diverges from the authors who use the embeddings[7] by Pyysalo et al. (2013a), induced on a combination of biomedical and general-domain data (PubMed, PMC and Wikipedia texts).

6 Results

We compare the performance of the baseline with the retrofitted embeddings models by measuring their precision (P), recall (R), and F_1-scores in text classification and relation extraction when used as input features.

For the text classification tasks, Tables 6 and 7 show the micro-averaged scores for the HOC and

[6] https://github.com/cambridgeltl/BioNLP-2016

[7] obtained from: http://bio.nlplab.org

the EXP tasks respectively. Each table shows the performance on document- and sentence-level classification (as columns) with different semantic lexicons (as rows).

For the relation classification task (CHEMPROT), Table 8 shows the micro-averaged scores. The best results are shown in bold and statistically significant scores are shown with an asterisk. All statistical tests are performed using a two-tailed t-test with $\alpha = 0.05$.

We first describe experiments measuring improvements from the retrofitting method, followed by comparisons against using different sets of lexicons during retrofitting.

6.1 Retrofitting

We use Equation 1 to retrofit word representations using linguistic constraints derived from verb lexicons. Overall, the retrofitted models show improvements in most tasks.

For text classification, the scores have improved in three out of the four cases. For the HOC task (Table 6) all retrofitted models outperform the baseline in F_1-score, which is largely attributed to a substantial improvement in recall (particularly for document-level classification, where there is a 15 point increase over the baseline). In total, five out of the twelve improved scores reported are also statistically significant.

The results for the EXP task (Table 7) are more mixed. At the document level, all retrofitted models achieve a slight F_1-score gain and half of the scores are significant. There is an improvement in recall at the cost of lower precision when compared to the baseline.

However, we can see that sentence-level classification is more difficult, due to the smaller amount of context information available. On the sentence level, the baseline seems to outperform all others, and only two out of six cases are significant. It indicates that the lexicons did not aid sentence-level classification in this particular task.

In relation classification, the word representation achieves the state-of-the-art result after incorporating our lexical information (34 classes). From Table 8, there is approximately a 1.5 point F_1-score increase over the baseline, and half of the improvements reported are significant. The results from both tasks suggest that the class-features provided by verb lexicons improve performance over the raw verb features.

6.2 Semantic lexicons

We compare the performance of our retro-fitted embeddings using both expanded clusters and the manually annotated clusters lexicon. The expanded clusters retrofitted embeddings outperform the original annotated clusters retrofitted embeddings in all evaluated tasks. This is likely due to the larger size of the expanded clusters in comparison to annotated clusters (Table 1), thus providing features for more verbs.

Lexical resources can be useful for NLP tasks for their abilities to capture generalizations about a range of linguistic properties; however, the degree of generalization needed may vary from task to task. When experimenting with retrofitting with different levels of verb classes, we observe a notable difference (1–2 points in F_1-score) between models retrofitted with the coarse-grained level of 16 classes and the fine-grained level of 50 classes.

For document-level text classification in both datasets (Tables 6 and 7), models appear to benefit from a finer-grained classification of 50 classes; on the sentence level a medium level of generalization (34 classes) seems optimal. The best result for relation classification (Table 8) is also obtained with 34 classes.

7 Discussion

The task-based evaluations suggest that verb clusters and a verb-optimized representation, can be a useful resource to support biomedical NLP tasks. In text classification, it has been observed that the occurrence patterns of verbs can be "topic-related" and certain set of verbs frequently appear within a specific topic of documents (Doan et al., 2009; Hatzivassiloglou and Weng, 2002; Sekimizu et al., 1998). Regarding this, expanded clusters appears to have captured some of these topic-related properties. On the HOC dataset, we note that some frequent verbs (such as "proliferate" and "grow") appearing in documents relating to the topic *Sustaining proliferative signaling* also share the same classes in our automatically-created lexicon. Similarly, for exposure assessment documents describing air monitoring data in EXP, we can frequently see member verbs such as "inhale" and "breathe" in the *proceed* class.

Entities–relations described in the biomedical literature are often expressed in a predicative form where a trigger word (most commonly a verb) connects two or more entities; here a range of

Lexicon	Document classification			Sentence classification		
	P	R	F_1	P	R	F_1
No lexicon & SOTA	77.8	51.7	62.1	56.8	30.7	39.9
Annotated clusters						
16-classes	75.1	56.4	64.8	47.1	34.6	39.9
34-classes	74.2	56.6	64.3	48.4	35.5	41.0
50-classes	74.9	59.2	66.2	48.4	35.2	*40.7
Expanded clusters						
16-classes	75.5	64.4	*69.5	45.2	36.5	*40.4
34-classes	74.3	63.5	*68.5	52.7	35.6	**42.5**
50-classes	73.9	66.1	***69.8**	50.9	34.7	41.3

Table 6: Performance results for the Hallmarks of Cancer task (HOC) when different sets of lexicons are used for retrofitting the baseline model. Baseline denotes a skip-gram model generated with our optimized training settings. Scores are adopted from Baker and Korhonen (2017). All figures are micro-averages expressed as percentages (Bold: the best score, *: statistically significant).

Lexicon	Document classification			Sentence classification		
	P	R	F_1	P	R	F_1
No lexicon & SOTA	89.5	87.1	88.3	66.2	62.8	**64.5**
Annotated clusters						
16-classes	88.9	87.7	*88.3	67.1	58.9	62.7
34-classes	89.4	87.8	*88.6	67.2	58.2	*62.4
50-classes	88.9	88.7	88.8	65.6	55.7	60.3
Expanded clusters						
16-classes	89.2	87.9	88.5	66.7	60.0	63.2
34-classes	88.7	88.9	*88.8	67.3	58.7	62.7
50-classes	88.6	89.1	**88.9**	67.5	58.6	*62.7

Table 7: Performance results for the Chemical Exposure Assessment task (EXP). Baseline denotes a skip-gram model generated with our optimized training settings. The "No lexicon" scores are from Baker and Korhonen (2017). All figures are micro-averages expressed as percentages. (Bold: the best score, *: statistically significant).

verbs can be used to describe similar relations. Understanding the commonalities shared among individual verbs helps NLP systems to identify the particular type of relation the text is describing. Consider as an example the *suppress* class in our verb lexicons. It captures the fact that its members are similar in terms of syntax and semantics, and they can be used to make similar statements which describe similar events. In CHEMPROT, member verbs in the *suppress* class such as "suppress" and "inhibit" can often be found in sentences depicting the *down-regulation* relation between chemicals and proteins.

For many NLP applications, lexical classes are useful for their ability to capture generalizations about a range of linguistic properties: by retrofitting word representations to lexical resources, semantically similar verbs (*i.e.* member verbs within the same lexical class) like "suppress" and "inhibit" will be pulled together in the vector space, whereas verbs like "collect" and "examine" will not. Consequently, this allows NLP systems to generalize away from individual verbs, alleviating the data sparseness problem of representing each verb in the corpus individually. The lexical classes provide different levels of generalization to support tasks of various needs, from the coarse-grained level of 16 classes to a fine-grained level of 50. A notable performance difference is observed when we evaluate models retrofitted with different levels of verb classes. Among all three classes, we observe a larger improvement over models at the finer-grained levels of 34 or 50 classes, which reveal that finer-grained levels of verb semantic distinction seem more contributive in our assessed tasks.

Lexicon	P	R	F_1
No lexicon	76.9	63.5	*69.5
SOTA	75.1	65.1	69.7
Annotated clusters			
16-classes	76.5	64.6	70.1
34-classes	78.2	63.8	*70.3
50-classes	76.5	65.0	*70.3
Expanded clusters			
16-classes	76.3	65.2	70.3
34-classes	77.5	65.6	**71.0**
50-classes	76.2	65.9	*70.7

Table 8: Performance results for the Chemical-Protein Interaction (CHEMPROT) when different sets of lexicons are used for retrofitting the baseline model. Baseline denotes a skip-gram model generated with our optimized training settings. SOTA denotes the state-of-the-art result reported by Björne and Salakoski (2018) using the embeddings by Pyysalo et al. (2013a). All figures are micro-averages expressed as percentages. (Bold: the best score for the task, *: statistically significant).

8 Conclusions

Many core NLP tasks and applications in the biomedical domain such as relation and event extraction, text classification, and text mining may benefit from accurate embedded representation of verbs.

Verb semantic classes capture generalizations about a range of linguistic properties, by retrofitting embedded word representations to semantic verb classes, semantically similar verbs (*i.e.* verbs that are members of the same lexical class) are pulled together in the vector space. Consequently, this allows NLP systems to generalize away from individual verbs, reducing the problem of data sparseness in representing less frequent verbs.

The key contribution of this work is to show that by using semantic classes for verbs (such as those provided by both the annotated and expanded clusters) we can improve the downstream performance on several tasks in the biomedical domain by aligning word embeddings according to semantic verb classes.

This is achieved by a post-processing retrofitting procedure, using a standard "off-the-shelf" method, by running belief propagation on a graph constructed from lexicon-derived relational information to update word vectors. It can be applied to any pretrained word embedding vectors.

We applied two lexicons of semantic verb clusters to two sets of widely used pretrained embedding vectors in BioNLP on several downstream tasks: two text classification tasks (the Hallmarks of Cancer, and Chemical Exposure Assessment) with both document and sentence classification, as well as a relation extraction task (CHEMPROT). We used a standard "off-the-shelf" retrofitting algorithm to obtain improved embeddings, and we feed the retrofitted representation to the current state-of-the-art models for their respective tasks. We controlled the experimental setup by using the same model implementation, as well as the same training, development and test data folds.

The results show that using verb clusters to retrofit embeddings, we achieved new state-of-the-art performance in the evaluated downstream tasks (with statistically significant scores); the only exception being sentence level classification for the Chemical Exposure Assessment task (however we do improve SOTA in document level classification for the same task). We also note a performance difference when retrofitting with different levels of verb classes, where we see a larger improvement when using finer-grained levels of verb semantic classes (30 or 50 classes), which seem more contributive.

For future work, we will further investigate the possibility of using verb lexicons for retrofitting new generations of word representation models such as contextualized embeddings; we will further evaluate on other downstream biomedical tasks, for instance event and pathway extraction and medical question answering.

Acknowledgement

This work is supported by the Medical Research Council [grant number MR/M013049/1], the ERC Consolidator Grant LEXICAL [grant number: 648909], the ESRC Doctoral Fellowship [grant number: ES/J500033/1] and the Defense Advanced Research Projects Agency [DARPA 15-18-CwC-FP-032].

We would like to thank our reviewers for their constructive feedback. We are very grateful to Tyler Griffiths for helping with proofreading and typesetting this paper.

References

Simon Baker, Imran Ali, Ilona Silins, Sampo Pyysalo, Yufan Guo, Johan Högberg, Ulla Stenius, and Anna Korhonen. 2017. Cancer hallmarks analytics tool (chat): a text mining approach to organize and evaluate scientific literature on cancer. *Bioinformatics*, 33(24):3973–3981.

Simon Baker and Anna Korhonen. 2017. Initializing neural networks for hierarchical multi-label text classification. *BioNLP 2017*, pages 307–315.

Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan Högberg, Ulla Stenius, and Anna Korhonen. 2015. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, 32(3):432–440.

Jari Björne. 2014. *Biomedical Event Extraction with Machine Learning*. Ph.D. thesis, University of Turku.

Jari Björne and Tapio Salakoski. 2018. Biomedical event extraction using convolutional neural networks and dependency parsing. In *Proceedings of the BioNLP 2018 workshop*, pages 98–108.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 173–180. Association for Computational Linguistics.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical nlp. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 166–174.

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

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

K Bretonnel Cohen, Martha Palmer, and Lawrence Hunter. 2008. Nominalization and alternations in biomedical language. *PloS one*, 3(9):e3158.

Jessica Cox, Corey A Harper, and Anita de Waard. 2017. Optimized machine learning methods predict discourse segment type in biological research articles. In *Semantics, Analytics, Visualization*, pages 95–109. Springer.

Son Doan, Ai Kawazoe, Mike Conway, and Nigel Collier. 2009. Towards role-based filtering of disease outbreak reports. *Journal of Biomedical Informatics*, 42(5):773–780.

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

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

Vasileios Hatzivassiloglou and Wubin Weng. 2002. Learning anchor verbs for biological interaction patterns from published text articles. *International Journal of Medical Informatics*, 67(1-3):19–32.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Karin Kipper, Hoa Trang Dang, Martha Palmer, et al. 2000. Class-based construction of a verb lexicon.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of english verbs. *Language Resources and Evaluation*, 42(1):21–40.

Anna Korhonen, Yuval Krymolowski, and Nigel Collier. 2006. Automatic classification of verbs in biomedical texts. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 345–352. Association for Computational Linguistics.

Martin Krallinger, Obdulia Rabal, Saber A Akhondi, et al. 2017. Overview of the biocreative vi chemical-protein interaction track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*, volume 1, pages 141–146.

Kristin Larsson, Simon Baker, Ilona Silins, Yufan Guo, Ulla Stenius, Anna Korhonen, and Marika Berglund. 2017. Text mining for improved exposure assessment. *PloS one*, 12(3):e0173132.

Ben Lengerich, Andrew Maas, and Christopher Potts. 2018. Retrofitting distributional embeddings to knowledge graphs with functional relations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2423–2436. Association for Computational Linguistics.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Bill MacCartney, Christopher D Manning, and MC de Marneffe. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings LREC*.

G Suryanarayanan Mahalakshmi. 2015. Content-based information retrieval by named entity recognition and verb semantic role labelling. *Journal of universal computer science*, 21(13):1830.

David Mcclosky. 2010. *Any Domain Parsing: Automatic Domain Adaptation for Natural Language Parsing*. Ph.D. thesis, Brown University, Providence, RI, USA. AAI3430199.

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

Sampo Pyysalo, Tomoko Ohta, and Sophia Ananiadou. 2013b. Overview of the cancer genetics (cg) task of bionlp shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 58–66.

CJ Rupp, Paul Thompson, William Black, John McNaught, and Sophia Ananiadou. 2010. A specialised verb lexicon as the basis of fact extraction in the biomedical domain. *Proceedings of Verb 2010*, page 188.

Karin Kipper Schuler. 2005. Verbnet: A broad-coverage, comprehensive verb lexicon.

Takeshi Sekimizu, Hyun S Park, and Jun'ichi Tsujii. 1998. Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts. *Genome informatics*, 9:62–71.

Abhishek Sharma, Rajesh Swaminathan, and Hui Yang. 2010. A verb-centric approach for relationship extraction in biomedical text. In *2010 IEEE Fourth International Conference on Semantic Computing*, pages 377–385. IEEE.

RA Weinberg and Douglas Hanahan. 2000. The hallmarks of cancer. *Cell*, 100(1):57–70.

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

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

Zhiguo Yu, Byron C Wallace, Todd Johnson, and Trevor Cohen. 2017. Retrofitting concept vector representations of medical concepts to improve estimates of semantic similarity and relatedness. *arXiv preprint arXiv:1709.07357*.