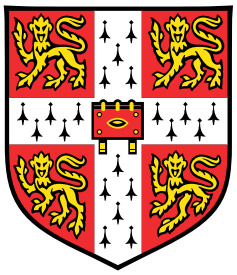
Neural Word Representations for Biomedical NLP



Chiu Hon Wing

Department of Theoretical and Applied Linguistics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Fitzwilliam College

June 2019

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 80,000 words including appendices, bibliography, footnotes, tables and equations.

Chiu Hon Wing
June 2019

Acknowledgements

I would first like to thank my supervisor, Prof. Anna Korhonen, for her patience, motivation and continuous support of my doctoral study and related research. She consistently allowed me to work independently, but steered me in the right direction whenever she thought I needed it. Her guidance helped me in all the time of research and writing of this thesis. Also, I want to thank her for all of the collaboration opportunities I was given to conduct my research. I could not have imagined having a better advisor and mentor for my study.

I would also like to thank my second supervisor, Dr Nigel Collier, who provided constructive feedback and detailed comments on my research with his expertise in biomedicine.

Besides my supervisors, I would like to thank Dr. Sampo Pyysalo, for his insightful comments and encouragement, but also for the hard questions which incented me to widen my research from various perspectives.

I would also like to acknowledge Dr Mohammad Taher Pilehvar, Dr Ivan Vulić and Dr Simon Baker, I have been extremely lucky to have been directly or indirectly collaborate with them, and I am gratefully indebted to their very valuable comments throughout my doctoral study and writing this thesis.

With a special mention to Gamal Crichton, Daniela Gerz, Olga Majewska, Edoardo Ponti and Milan Gritta for their wonderful collaboration. They supported me greatly and were always willing to help me. They have given me the best memories and experiences in Cambridge, and I am going to miss our interesting and long-lasting chats. Also to everyone in the LTL lab, it was great sharing laboratory with all of you during last four years.

I would also like to thank the experts both from Cambridge and Karolinska Institutet who were involved in the validation survey for this thesis. Without their passionate participation and input, the validation survey could not have been successfully conducted.

Finally, I must express my very profound gratitude to my parents and to my partner, Yan, for providing me with unfailing support and continuous encouragement throughout my years of study and my life. They are always there for me. This accomplishment would not have been possible without them. Thank you.

Abstract

Word representations are mathematical objects which capture the semantic and syntactic properties of words in a way that is interpretable by machines. Recently, the encoding of word properties into a low-dimensional vector space using neural networks has become popular. Neural representations are now used as the main input to Natural Language Processing (NLP) applications and in most areas of NLP, achieving cutting-edge results.

Our work extends the usefulness of neural representations, with a particular emphasis on the biomedical domain which is linguistically highly challenging. We focus on three directions: first, we present a comprehensive study on how the quality of the representation model varies according to its training parameters. For this, we implement a set of well-established models with different training settings regarding the size of input corpora, model architectures and hyper-parameters, and evaluate them thoroughly using the standard methods. Our best model significantly outperforms the baseline one, demonstrating the high impact of training parameters and the necessity of their optimization. The study provides an important reference for researchers using neural representations for biomedical NLP. Second, we introduce two novel datasets for evaluating noun and verb representations in biomedicine. These datasets are designed to be consistent with those available for mainstream NLP. They enable, for the first time, evaluation of verb representations in the domain. Last, we propose a neural approach to facilitate the development of a VerbNet-Style classification in biomedicine: we start from a small manual classification of biomedical verbs and apply a state-of-the-art neural representation model, developed explicitly for verb optimization, to expand that classification with new members. Evaluation of the resulting resource shows promising results when representation learning is performed using verb-related contexts. Additionally, our human- and task-based evaluations reveal that the automatically-created resource is highly accurate, suggesting that our method can be used to facilitate cost-effective development of verb resources in biomedicine.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

The performance of Natural Language Processing (NLP) systems depends heavily on the choice of data representation. Hence, there is active research in the NLP community regarding how to best represent data in terms of features that can support downstream applications. In current NLP, such techniques are commonly based on *Representation learning*.

1.1 Word representation learning

Representation learning, when applied to textual data, generates word representations which capture the linguistic properties of words in a mathematical form (e.g. vectors). Each vector dimension corresponds to a feature that might have a semantic or syntactical interpretation [Turian et al., 2010]. Early studies mostly use human experts to propose a set of representative features for the data, which is expensive to obtain. Recently, the unsupervised approach, which encodes word meanings into a low-dimensional space using neural networks has been suggested as an alternative [Bengio et al., 2003]. Such an approach, namely *neural embeddings* or *word embeddings*, represents each word as a dense vector of real numbers, where synonyms appear as neighbours in vector space. It can learn features unsupervisedly from large unlabelled corpora.

Despite word embeddings' usefulness, most studies adopt a unified learning approach towards different word-types (e.g. nouns and verbs). Since individual word-type often has certain unique linguistic properties, a single learning approach generally cannot capture the semantics of all word-types. Hence, there is a need to fine-tune representation learning algorithms so that they can effectively learn the properties for individual word-types (e.g. verbs).

While word embeddings have been shown to be beneficial in recent work, most of these studies are carried out with general-domain texts and evaluation datasets, and their results

do not necessarily apply to texts in other domains (e.g. biomedicine) that are linguistically distinct from general English. To get the maximal benefit from using word embeddings for biomedical NLP tasks, they need to be induced and evaluated using in-domain texts.

1.2 Biomedical NLP

The application of NLP methods to the biomedical domain has long been active because automated text processing systems need to handle the exponential growth of in-domain literature. The performance of these methods depends heavily on the choice of data representation. Thus, much of the efforts in biomedical NLP focus on how to best represent biomedical texts in terms of features that can support effective applications.

The quality of word representations relates closely to their training settings, including the sizes of input corpora, model architectures and hyper-parameters. In recent years, a number of novel representation learning models have been proposed and they have shown to be useful in supporting a range of NLP tasks. However, only few studies compare among existing models under different training settings. Hence, the impact level of a particular model's training parameters towards its quality is still uncertain.

Another critical concern stems from the means to measure the quality of representation models. Evaluation methods are broadly categorized into two types: the intrinsic and extrinsic evaluations. A typical intrinsic evaluation is the word similarity task. Given a list of word pairs rated with different degrees of similarity by annotators, the task compares the similarity-ranking produced by humans and representation models. The best model is deemed to be the one that gives the closest match to humans' ground-truth. On the other hand, extrinsic evaluation refers to the task-based evaluation, where the quality of a model is measured by how well it performs when used as a feature for NLP tasks. Since intrinsic evaluation is easy to implement, it is commonly used as a proxy measurement before a model is deployed in NLP applications. Consequently, intrinsic evaluation is expected, to an extent, to reflect how individual models perform in extrinsic tasks, but this presumption has not been verified in the research community. Furthermore, although the verb is vital to the meaning interpretation of biomedical language, the field currently lacks intrinsic evaluations for verb representations.

1.3 Verb classification

The verb forms an indispensable part of sentences, especially biomedical verbs that describe the actions between entities because it frames the types, numbers, and relations of participants in the event it describes. To help biomedical NLP researchers in identifying the particular

type of relation between entities described in the text, it is essential to have resources that gather the syntactic and semantic information about verbs. An example in the general domain is VerbNet [Kipper-Schuler, 2005]. It contains information about different linguistic properties of a set of verbs and the relationships among them. Verbs are grouped into a finite number of semantic classes, and members in the same class share similar properties (e.g. syntax). Though VerbNet has shown to be useful to support a wide range of NLP tasks, including semantic role labelling and other text mining applications [Lippincott et al., 2013; Rimell et al., 2013; Schmitz et al., 2012], it is created manually by linguists to describe verb properties in general English, so it provides limited coverage of verbs for other domains (e.g. biomedicine) where the linguistic properties are different from general English. It is foreseeable that resources of a similar nature can benefit biomedical NLP, yet the vast majority of in-domain resources merely cover noun concepts, and only a few small-scale resources for verbs can be found. Also, most resources are manually created and difficult to extend.

There is a necessity for the use of NLP techniques to automate the construction of lexicons. In this regard, automatic lexical acquisition refers to the automatic or semi-automatic process of learning lexical resources from unstructured text. Recent studies demonstrate that VerbNet-style lexicon can be acquired unsupervisedly from general and biomedical texts through automatic verb classification [Joanis et al., 2008; Kawahara et al., 2014b; Korhonen et al., 2006, 2008; Peterson et al., 2016; Sun, 2013; Vlachos et al., 2009]. However, existing approaches rely heavily on time-consuming feature engineering processes to extract linguistic properties from corpora. In contrast, unsupervised neural embedding methods provide an alternative to learning word features from a large unlabelled text. They have shown to be effective for inducing verb lexicons in the general domain [Vulić et al., 2017]. In the biomedical domain, studies of similar nature are limited partly because there is currently a lack of in-domain (intrinsic) evaluation for verb representations.

In general, representation learning is an important technique in NLP because it efficiently extracts useful features from data. The quality and usefulness of word representations obtained can be further improved through various ways such as optimizing their training parameters and developing word-type specific learning approaches. Despite this, biomedical verb representations currently lack an intrinsic evaluation metric and other measures that can effectively reflect the performance of representation models in extrinsic tasks. Lastly, although verbs are essential, most existing lexicons in biomedicine only cover nouns and are manually created; therefore, to have verb lexicons together with methods that can automate lexical acquisition are crucial.

1.4 Our contributions

This thesis aims to extend the utility of word representations to a domain which has distinct language properties as compared to general English (i.e. biomedicine) as well as to a lexical task of specific word-types (i.e. verbs).

Our contributions include:

- We investigate how the state-of-the-art representation model (namely, the **word2vec** package, [Mikolov et al., 2013a]) developed for general English can be transferred to a new domain. We experiment with biomedical texts and explore how the sizes of input corpora, model architectures, and hyper-parameters individually affect the quality of neural representations, as measured with intrinsic and extrinsic evaluations in biomedicine. We observe that models perform notably better (5 points in F-score) in practical NLP tasks after fine-tuning for in-domain text. Our fine-tuned model also achieves the state-of-the-art score in another study [Baker et al., 2016].

- Existing intrinsic evaluation datasets in biomedicine measure the quality of noun representations only. Additionally, these benchmarks fail to reflect how well individual models perform in practical NLP tasks. Hence, we create two new evaluation datasets: Bio-SimVerb and Bio-SimLex, which facilitate evaluations of biomedical verb and noun representations. Compared with existing results in biomedicine, the evaluation results from our datasets correlate better with the performance of NLP tasks.

- We explore how unsupervised neural representations can be used to facilitate cost-effective lexical acquisition for biomedical verbs. In particular, we propose an approach that can automatically identify the set of contributive contexts for learning biomedical verb representations from large amounts of text without manual feature engineering. As evaluated with Bio-SimVerb (our proposed evaluation dataset), the representation model shows improvement when it is trained using only verb-related contexts. We then apply our verb-optimized model to a small manual classification of biomedical verbs and generate a large lexical class of biomedical verbs.

- We provide both qualitative and quantitative evaluations for our automatically-acquired lexical classes. We illustrate through human- and task-based evaluations that lexical classes, as induced by verb-optimized representations, are highly accurate. They include novel, valid member verbs and classes, and can provide useful features to tasks like topics and relations identification in scientific abstracts.

1.5 Thesis outline

The remainder of this thesis is structured into six chapters. The brief overview of each chapter is as followed:

- In chapter 2, we present the background context and related literature for word representation models and describe their evaluation metrics that are used in this thesis.
- In chapter 3, we implement the cutting-edge representation learning approach, performing both intrinsic and extrinsic evaluations, in order to investigate its optimal training settings (e.g. model architectures and hyper-parameters) for biomedical NLP tasks. We highlight some notable practices and settings that are worth noticing when training word representations for biomedical tasks. Most importantly, when we assess the context window size (one of the training parameters), we find that results from all existing intrinsic evaluation benchmarks in biomedicine fail to reflect how individual models perform in extrinsic tasks (referred to as intrinsic–extrinsic contradiction henceforth).
- In chapter 4, we further look into the intrinsic–extrinsic contradiction and investigate whether this issue is domain-dependent. We train a set of representation models on general-domain text using different context window sizes and performing both intrinsic and extrinsic evaluations using general-domain datasets. We have observed a similar contradiction for general-domain evaluations and models.
- In chapter 5, we describe the creation of two new intrinsic evaluation datasets: Bio-SimVerb and Bio-SimLex. They can be used to evaluate noun and verb representations in biomedicine. They are created to address the intrinsic–extrinsic contradiction which is identified previously in chapter 3 and chapter 4, and they allow evaluations for representation models of other word-types such as verbs in biomedicine.
- In chapter 6, we investigate how word representations can be optimized for capturing semantic properties of biomedical verbs. We take our optimized model to construct a verb lexicon. In addition to evaluating our lexicon against human judgments, we also perform the task-based evaluation on text classification and relation classification, demonstrating word representations, after performing verb-optimization, have the potential to induce type-specific resources which can be used to support NLP tasks in biomedicine.
- In chapter 7, we summarize the contributions of the thesis and provide directions for future work.

1.6 Relevant publications

A number of peer-reviewed publications were produced while conducting the work in this thesis. They are mentioned below, and a note is made to their relevant chapters.

1. **How to train good word embeddings for biomedical NLP [Chiu et al., 2016a].**

   We conduct a comprehensive study to investigate the optimal settings (the size of input corpora, model architectures and hyper-parameters) to train word representations for biomedical named entity recognition. We report a notable improvement in both intrinsic and extrinsic evaluation after the optimization, highlighting its importance when using representation models in practical biomedical tasks. Apart from noticing that a larger corpus does not necessarily guarantee a better result, we also reveal that one training parameter can lead to contradictory results between intrinsic and extrinsic evaluations (i.e. the intrinsic–extrinsic contradiction). *This publication is relevant to Chapter 3.*

2. **Intrinsic evaluation of word vectors fails to predict extrinsic performance [Chiu et al., 2016b].**

   After we observe the intrinsic–extrinsic contradiction in existing biomedical datasets, we further examine whether such issue is domain-specific and conduct experiment using general texts, finding evidence that the intrinsic–extrinsic contradiction is domain-independent. In experimenting, an exception intrinsic dataset in the general domain is found (SimLex-999, [Hill et al., 2015]). We hypothesize that its consistent result with extrinsic evaluation can be attributed to its unique design protocol which distinguishes between the concepts of synonymy and relatedness (e.g. *pill* and *tablet* v.s. *doctor* and *medicine*). We thus follow its design principle and develop two novel intrinsic datasets for biomedicine: Bio-SimLex and Bio-SimVerb. *This publication is relevant to Chapter 4.*

3. **Bio-SimVerb and Bio-SimLex: wide-coverage evaluation sets of word similarity in biomedicine [Chiu et al., 2018].**

   Given the intrinsic–extrinsic contradiction in existing biomedical evaluation standards, we create Bio-SimLex and Bio-SimVerb which evaluate noun and verb representations in biomedicine respectively. Both datasets show a more consistent intrinsic–extrinsic estimation as compared with all existing intrinsic datasets in biomedicine. Bio-SimVerb also symbolizes the first intrinsic evaluation for verb representations in biomedicine. *This publication is relevant to Chapter 5.*

4. **A Neural Classification Method for Supporting the Creation of BioVerbNet [Chiu et al., 2019].**

   We extend the usefulness of representation models by optimizing their learning for a particular word-type: verbs. While existing methods typically deploy a single learning approach for different word-types in corpora, we show that the quality of representation learning for verbs can be further improved by filtering out those syntactic contexts which are less contributive to verb semantics (e.g. noun modifiers). In view of a lack of large-scale verb lexicon in biomedicine, we further apply the verb-optimized representations to construct a large lexical class for biomedical verbs. Human validation by domain experts reveal that the resource, as induced by a verb-optimized representation model, is highly accurate, suggesting that it can facilitate cost-effective development of verb lexicons. *This publication is relevant to Chapter 6.*

Chapter 2

Background

This chapter provides the reader with background information relevant to our work. We first summarize the well-established algorithms and evaluation metrics that are used for representation learning, followed by a review on how they have been applied in biomedical Natural Language Processing (NLP). Then, we describe some lexical resources for NLP (mainly on verbs), and provide a survey on approaches used for lexical acquisition (both the manual and automatic ones), covering the data, features and models used in these approaches.

2.1 Representation learning

For many NLP applications, the choice of data representation is a critical aspect of achieving strong performance. So much of the effort in deploying NLP algorithms goes into *feature engineering* that can best represent linguistic features in data that can support downstream applications. Though useful, feature engineering suffers from several disadvantages. First, feature engineering can be time-consuming and often requires expert knowledge to produce informative features. Second, feature engineering is computationally demanding since it often requires mature NLP pipelines, such as part-of-speech (POS) tagging, dependency parsing, and named entity Recognition (NER), to extract features from the text. Last, the set of extracted features are often domain-dependent, which implies they may not be universally representative to the text in every domain.

To ease the use of feature engineering, one option is to automate the feature learning process. In this regard, *representation learning* refers to the set of techniques that automatically discovers and extracts useful features in data needed for classification and prediction tasks [Bengio et al., 2013]. Representation learning, when applied to textual data, generates the word representation which captures the linguistic features of words in vector form. Each word is associated with a finite-dimensional vector of real numbers (better known as word

vector), and each vector dimension corresponds to a feature which might have a semantic or syntactic meaning [Turian et al., 2010].

Representation learning is developed on the basis of the distributional hypothesis, which suggests that lexical items with similar distributions share similar meanings. More specifically, words that are used and occur in the same contexts tend to have similar meanings [Harris, 1954]. Thus, the core principle of word representation learning algorithms is to find distributionally similar words (e.g. words having similar co-occurrence counts in corpora) and assign them word representations that can map them to proximate regions in vector space. Since distributional information is largely available for many languages and can be extracted easily from large unannotated texts without depending on other NLP pipelines, the unsupervised learning of word representations using the distributional hypothesis has become widely popular.

Recent literature has suggested methods using neural networks to learn word representations [Bengio et al., 2003]. Thus far, neural representations have been widely-used to provide useful features to many successful NLP applications. Nevertheless, most existing models deploy a single learning algorithm and representational form for all words in the corpus; this disregards the individual word-type (nouns/verbs) and text-domain differences. Applying learning algorithms tailored to individual word-types is essential not only because each word-type often has some unique linguistic properties that are distinct from one another, but also learning algorithms can greatly extend the usefulness of word representations for applications of a particular word-type (e.g. building verb lexical resources). Besides, work in domain-NLP (e.g. biomedicine) has revealed that representation learning tends to be domain-dependent [Stenetorp et al., 2012], implying that representation models need to be learned from the in-domain text in order to obtain the maximal benefit for domain-NLP tasks. The current challenge, therefore, is to optimize these techniques for word-type (verbs) and domain-specific (biomedicine) applications. Next, we will describe some well-known models in representation learning, along with their evaluation methods.

2.1.1 Representation models

Building a vector representation of word semantics begins with extracting distributional information such as the co-occurrence frequencies between words from a text corpus. To illustrate, consider three sentences below:

1. I love chemistry.

2. I love maths.

3. I tolerate biology.

When one extracts co-occurrence frequencies at the sentence level, every word is said to be in the context of another word in the same sentence, thus the corpus can be represented in the following matrix form:

$$\mathbf{M} = \begin{array}{c} \\ I \\ love \\ Chemistry \\ Maths \\ tolerate \\ Biology \end{array} \begin{array}{c} \begin{array}{cccccc} I & love & Chemistry & Maths & tolerate & Biology \end{array} \\ \begin{pmatrix} 0 & 2 & 1 & 1 & 1 & 1 \\ 2 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

M represents the matrix of co-occurrence frequencies of words in the corpus. Each row of M is interpreted as the vector representation of the word corresponding to the row, and the columns are features that capture the word semantics. The similarity between words *Chemistry* and *Biology* can be computed as $cosine(\vec{Chemistry}, \vec{Biology}) = 0.5$.

Counts weighting

The co-occurrence matrix can be adjusted in a way that gives higher weight to less frequent co-occurrence pairs to signify the more informative context words out of the common ones. Other types of weights, including the Pointwise mutual formation (PMI) [Church and Hanks, 1990], have also been suggested in literature. PMI is given as:

$$\log \frac{P(x,y)}{P(x)P(y)} \tag{2.1}$$

$P(x,y)$ is the joint probability of word x and y, whereas $P(x)$ and $P(y)$ is the probability of word x and y individually appearing in the text. In addition to this, the co-occurrence matrix can also be designed to capture syntactic features of words by considering the context of a word to be words that are co-related to it by a syntactic dependency relation in the text [Baroni and Lenci, 2011; Lin, 1998; Padó and Lapata, 2007].

Neural embeddings

Because the co-occurrence matrix requires a high vector dimension (i.e. the number of columns in M) to represent every word-to-word co-occurrence in a large corpus, few methods

that obtain lower dimensional representations from a matrix have been proposed (e.g. Principal Component Analysis, PCA). What is becoming popular is encoding word semantics into a dense vector using a neural network. This method is commonly known as *Word embeddings* or *Neural embeddings*. Neural embeddings' learning algorithm functions much like a language modelling task, whose goal is to predict the next word (referred to as contexts henceforth) given the previous ones in a sentence. Each word is represented as a finite-dimensional vector of real numbers, and the objective is to maximize the joint probability of a word and its contexts in terms of word vectors using a feed-forward neural network. Word vectors are updated using back-propagation and gradient descent.

Continuous Bag-of-Words (CBOW) and Skip-gram

The CBOW and Skip-gram are two cutting-edge representation learning algorithms introduced by Mikolov et al. [2013a,b] as part of the **word2vec** tool. CBOW and Skip-gram have been shown to produce highly competitive neural embeddings in many intrinsic and extrinsic tasks [Baker et al., 2016; Pyysalo et al., 2013a; Rei et al., 2016; Tsvetkov et al., 2015], as compared to early models such as Random Indexing [Kanerva et al., 2000] and Latent Semantic Analysis [Landauer and Dumais, 1997], among others.

CBOW and Skip-gram learn word representations through a neural network, which is composed of an input layer, a fully connected hidden layer, and an output layer. The input layer size equals to the vocabulary size of the corpus, and each word is represented as a one-hot vector. (i.e. a vector of size $|V|$ where one dimension of the vector is set to 1 to indicate a word, while other dimensions are set to 0). The hidden layer corresponds to the dimensions of the output word vectors. If a corpus consists of $|V|$ words and D is the dimension of these word vectors, then the hidden layer will be a matrix of size $V \times D$, where each row corresponds to a word (as illustrated in Fig 2.1 and Fig 2.2). The hidden layer output is essentially the product of the hidden layer weight matrix (which are the learned representations). The size of the hidden layer is a hyper-parameter pre-defined by users. While a higher dimension tends to capture better word representations, their training produces a larger word representation matrix and is more computationally costly [Mikolov et al., 2013c].

CBOW and Skip-gram have the objectives of maximizing the probability of an individual word given its contexts: $P(w_t|c(w_t)); w \in V$, where w_t refers to the root word (i.e. the target word to be trained), V is the vocabulary of the corpus, and $c(w_t)$ is the set of context words that surround the root. The size of the context window defines the range of words to be included as the context of a root word, which again, is a hyper-parameter pre-defined by users. For instance, a window size of 2 takes two words before and after a root word as its

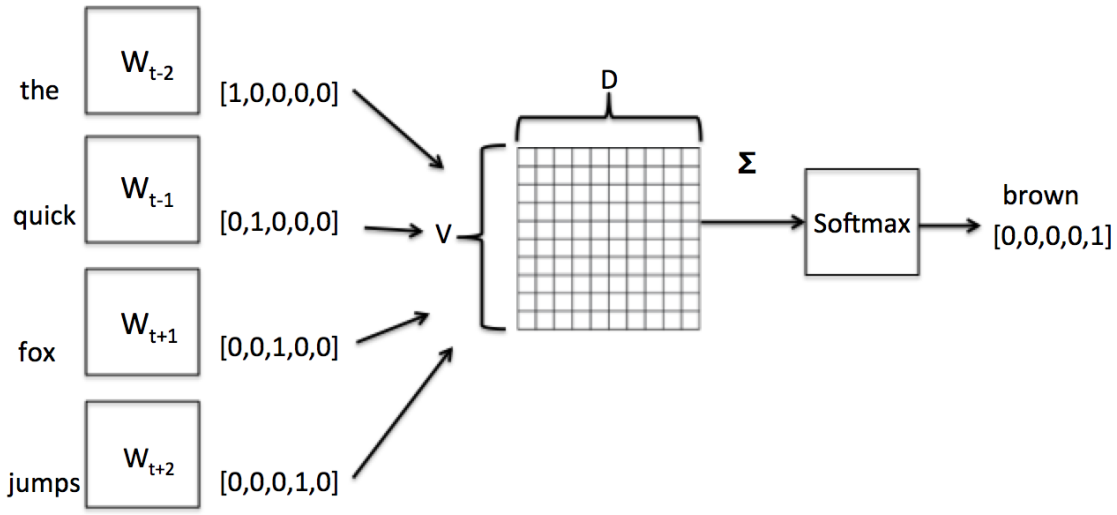


Fig. 2.1 An illustration of the CBOW model with window size 2. The model is predicting the root word 'brown' given the context 'the quick fox jumps'. V is the total words in the corpus, and D is the dimension of these word vectors. The symbol $\sum$ implies the average of the input context word $(c(w_t))$ vectors multiplied by the hidden layer weights. The Softmax function estimates a probability distribution over all words in the vocabulary.

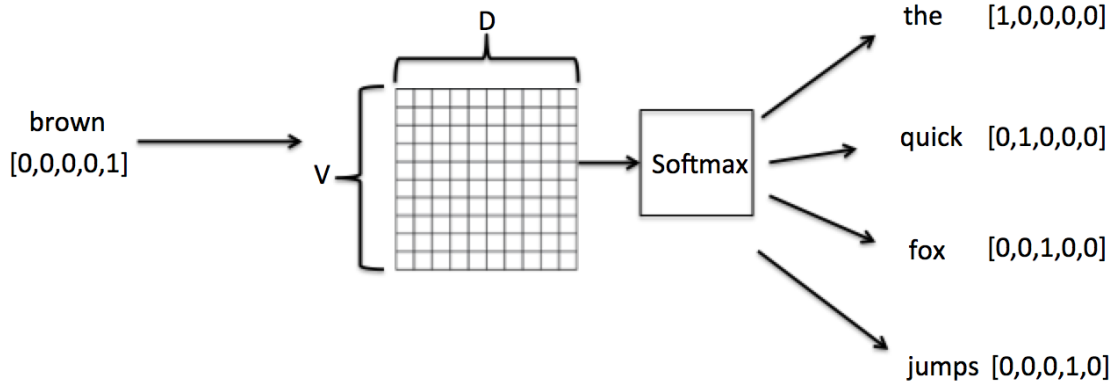


Fig. 2.2 An illustration of the Skip-gram model with window size 2. The model is predicting the root word 'brown' given the context 'the'. Every word-context pair will be trained individually. V is the total words in the corpus, and D is the dimension of these word vectors. The Softmax function estimates a probability distribution over all words in the vocabulary.

contexts for training. The window size is an important hyper-parameter in representation learning models because it controls the number of words to be considered as the context for representing an individual word. It may need a wider window when training on the text that is full of long sentences containing complex clausal structures (e.g. biomedical literature). Additionally, it has been shown that window size of a model influences the types of word semantics it captures: a larger window size emphasizes the learning of topic similarity between words, while a narrow context window leads the representation learning to primarily capture the word function [Turney, 2012].

A key difference between the trainings of CBOW and Skip-gram is the differentiated ways of denoting the context words (i.e. $c(w_t)$). In CBOW, context is denoted as the average of word vectors $\vec{c}_i$ within the window (size = i), which is calculated as followed:

$$c(w_t) = \frac{1}{|c(w_t)|}\Big(\sum_{c_i \in c(w_t)} \vec{c}_i\Big)^\top \tag{2.2}$$

In contrast, Skip-gram considers each context word in a window as a distinct vector, which is calculated as:

$$c(w_t) = \Big(\vec{c}_i\Big)^\top \tag{2.3}$$

Consequently, the output layer generates a probability value for the root word. This is done by converting the activation values output by the hidden layer into probabilities using the Softmax function as follows [1]:

$$P(w_t|c(w_t)) = \frac{\exp(c(\vec{w}_t)^\top \cdot \vec{w}_t)}{\sum_{v_i \in V} \exp(c(\vec{w}_t)^\top \cdot \vec{v}_i)} \tag{2.4}$$

The architectures of CBOW and Skip-gram share similar training parameters (e.g. context window size and vector dimension). Nevertheless, Skip-gram individually maps every word-context pair within a context window, making it intractable when used with a large amount of training data. Thus, its approximation counterpart – CBOW is introduced. It only estimates the probability of each root word with the average of contexts within the window. Other approximation techniques, such as the negative sampling and the sub-sampling [2], are also introduced as user-defined parameters in the word2vec package. These parameters control the number of training examples and facilitate the effective Skip-gram training in a large corpus. However, it is still uncertain how these training parameters influence the quality of the learned model.

[1] **Exp** stands for the Exponential function

[2] The description of the parameters is provided in Section 3.3.3

Context feature

In CBOW and Skip-gram, the representation of a word is learned by predicting all its neighbouring words within a window (contexts), assuming all contexts are useful. However, some contexts (e.g. stop words) co-occur with many words and thus are less informative features for distinguishing one word from another. To illustrate, when considering the similarity between the words *doctor* and *nurse*, the context *hospital* is more indicative than the contexts *have* or *like*. Additionally, some contexts that are useful for the representation learning of one word-type (e.g. nouns) may be uninformative for another one. For instance, a noun pre-modifier may be useful for learning noun representations but not verb representations. Consequently, other sources of contexts, such as the dependency relation between words, have been suggested in the literature. Such contextual features are shown to yield word vectors that follow functional similarity instead of the usual topical similarity [Turney, 2012]. In contrast, recent studies have highlighted the importance of developing learning algorithms, as well as exploring contexts for individual word-types [Schwartz et al., 2015; Vulić et al., 2017]. For example, Schwartz et al. [2015] expressed that symmetric patterns (e.g. x *and* y) yield significant improvements in learning representations for adjectives and verbs, while the traditional bag-of-words contexts are still the optimal choice for noun representation learning. Such word-type specific optimization can greatly extend the usefulness of representation models for tasks relating to a particular word-type (e.g. automatic lexical acquisition for verbs).

2.1.2 Evaluations for word representations

Two types of evaluations are commonly used to measure the quality of representation models: intrinsic and extrinsic evaluations. The typical intrinsic evaluation is word similarity, which consists of a list of word pairs which have been rated by humans with different degrees of similarity. Every rating measures the similarity between two words as perceived by a human; these words are rated on a scale of 1-10 (or any other scale provided individually for every dataset). These ratings are then aggregated across all raters to obtain an average measure of similarity for each word pair. A higher rating implies a more similar pair (e.g. *quick/rapid*: 8.75, *word/dictionary*: 3.68). To measure the intrinsic quality of a model, the researchers compute the cosine similarity of these word pairs by their corresponding vector representation. Then they measure the Spearman's rank correlation coefficient between the similarity-ranking produced by humans and models. The quality of a model is determined by the proximity between its similarity-ranking and the human ranking.

Conversely, in extrinsic evaluation, the quality of a model is estimated by how well it performs in NLP tasks. The most common performance metric is accuracy (A), which measures the proportion of instances that a classifier predicted the labels correctly:

$$A = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.5}$$

TP stand for the true positives, FP are the false positives, TN are the true negatives, FN are the false negative instances in the evaluation set.

The accuracy metric is sensitive to the size of datasets. If negative examples dominate a dataset, a trained classifier may be well-acquainted to classify negative examples but not the positive ones. The accuracy metric will fail to reflect such a scenario because the number of TN is still high even if TP is poorly predicted by the classier.

$Precision(P)$ and $recall(R)$ are suggested to address this issue. *Precision* measures the proportion of the positively classified instances that are correctly predicted by the classifier, whereas *recall* measures the proportion of positive instances in the data that the classifier is predict correctly. They are defined as followed:

$$P = \frac{TP}{TP + FP} \tag{2.6}$$

$$R = \frac{TP}{TP + FN} \tag{2.7}$$

$Precision(P)$ and $recall(R)$ are often combined using the harmonic mean, which is known as the $F_1 score$:

$$F_1 = \frac{2PR}{P + R} \tag{2.8}$$

The '1' in F_1 implies that both recall and precision are weighted equally. It is possible to select different weights like $F_{0.5}$ which give more weight to recall. In this work, all mentions of F-score refers to the standard F_1.

Evaluation resources in the general domain

For intrinsic evaluation, many word similarity datasets with different characteristics are created in the general domain. The sizes of these datasets vary largely, with the smaller datasets containing only 30 word pairs (MC-30) [Miller and Charles, 1991] to the larger datasets containing 3,000 word pairs (MEM-3000) [Bruni et al., 2012]. Some datasets are designed to evaluate the representations of a specific word-type likes nouns (RG-65) [Rubenstein and Goodenough, 1965] or verbs (YP-130) [Yang and Powers, 2006]. Recently, literature has

shown that the word pairs in most datasets are assessing both similarity and relatedness instead of exclusively assessing word similarity [Hill et al., 2015]. For example, *company* and *stock* are rated more similar than *train* and *car*, even though *train* and *car* share more common attributes (e.g. wheels and windows) than *company* and *stock* which are loosely related. Since the notion of word similarity is subjective and it is commonly confused with relatedness, the similarity and relatedness are suggested to be rated separately [Faruqui et al., 2016]. Subsequently, two word similarity datasets: SimLex-999 [Hill et al., 2015] and SimVerb-3500 [Gerz et al., 2016] are created to model attributional similarity rather than relatedness. The former contains a balanced set of noun, verb and adjective pairs, whereas the latter focuses on the evaluation of verb semantics.

Word similarity evaluation is easy to implement, and it can quickly assess how well the notion of word similarity according to humans is captured by the word representations. On the other hand, models can also be evaluated by their utilities as features in extrinsic tasks, such as text classification. With a lack of standardized extrinsic evaluation methods, intrinsic evaluation is therefore deemed to be a surrogate that provides quick estimation for models. Intuitively, models that can capture word similarity might perform well on tasks that require a notion of explicit semantic similarity between words like named entity recognition (NER). However, there are no empirical studies of how intrinsic evaluation is representative of the performance of models in their extrinsic tasks.

Recently, neural embeddings have been consistently proven to be useful in many NLP applications like NER and automatic lexical acquisition [Ma and Hovy, 2016; Vulić et al., 2017], yet most studies only involve general-domain texts and evaluation datasets, and their results do not necessarily apply to domain-NLP tasks (e.g biomedicine). For this, models need to be trained and evaluated with in-domain datasets for optimal performance. This would require resources such as scientific literature, evaluation datasets, lexicons, and terminologies specific to the domain.

2.2 Biomedical NLP

The application of NLP methods to biomedicine has become increasingly popular due to the need for techniques to automatically process information in the growing amount of scientific literature. This section describes some basic resources and tools used in biomedical NLP.

2.2.1 Scientific literature and representation models

Scientific literature serves as one of the main resources for many biomedical NLP applications. It is often available in the form of a large-scale database. Some examples include the PubMed abstracts and the PubMed Central open-access (referred to as **PubMed** and **PMC** henceforth). PubMed and PMC are the abstracts and the free full-text archive of biomedical and life sciences journal literature respectively, as maintained by the US National Library of Medicine. The rich literature constitutes an unannotated corpus of 5.5 billion tokens, covering the entire available biomedical scientific literature and forming a representative corpus of the domain. Thus, the topics on how to extract useful features from these unannotated texts for NLP tasks have been actively studied in the biomedical NLP community [Kosmopoulos et al., 2015; Stenetorp et al., 2012]. A number of representation models have been considered (details in Table 2.1). For example, Pyysalo et al. [2013a] studied the **word2vec** tool and found that Skip-gram produced highly competitive word representations in many intrinsic and extrinsic tasks, as compared to previous models such as Random Indexing and Latent Semantic Analysis, among others. Additionally, Muneeb et al. [2015] showed that Skip-gram outperformed other neural models such as Global Vectors (GloVe) [Pennington et al., 2014] on word similarity tasks.

Because the linguistic properties in the biomedical text differs significantly from general English (e.g. it is commonly written in long sentences containing a complex clausal structure and full of terminologies and acronyms), it is difficult to directly use models that are trained with general text for biomedical NLP. Hence, there is active research on how to fine-tune generic representation learning methods, particularly their training settings, to better adapt with the biomedical data for optimal performance. For example, Stenetorp et al. [2012] studied how the sizes and scopes of corpora affected the performance of various representation methods, including the Brown clusters and the Hierarchical Log-Bilinear embeddings (HLBL). As evaluated on three biomedical-NER datasets (AnatEM) [Ohta et al., 2012], BC2GM [Smith et al., 2008] and NCBID [Doğan and Lu, 2012]), they reported that models trained on larger in-domain corpora showed greater and more consistent benefits as compared with the ones induced on general texts. Although they evaluated the qualities across various representation models, they also highlighted the importance of assessing the effects of individual training parameters (e.g. corpus size and hyper-parameter values) for a single model, which is currently lacking in the field.

Word Representation	Descriptions
Brown clusters [Brown et al., 1992]	It is a hierarchical, bigram-based clustering algorithm. It introduces a binary tree on top of the corpora, refining information about the similarity of words with each branch. In the resulting representation, each word is assigned in a cluster that leads from the root of that tree to the leaf that the word is assigned to.
Google N-gram clusters [Lin et al., 2010]	Lin et al. [2010] introduced a new N-gram corpus from web-crawled data. Each word is assigned to clusters which are derived from K-means with distance defined by the dot product of vectors containing mutual information between a word and each of its context words.
Random Indexing [Kanerva et al., 2000]	Random indexing is a method for constructing a semantic word vector model in an incremental manner. First, every word is assigned an one-hot vector with all elements equal to zero, except for a small number of randomly distributed +1 and -1 values. The vector-space representation of a given word is then obtained by summing up the one-hot vectors of all words in all its context windows in the corpus
HLBL embeddings [Mnih and Hinton, 2009]	The Hierarchical Log-Bilinear embeddings is a distributed word representation. It is low dimensional, real-valued vectors with mostly non-zero components. The representation is induced using neural network-like language models. The HLBL embeddings is composed by condensing all model representations for all contexts of a given word.
CW embeddings [Collobert and Weston, 2008]	The Collobert and Weston embeddings is inferred directly as part of a neural network aimed at solving a particular NLP task (e.g POS tagging).
BioASQ embeddings [Kosmopoulos et al., 2015]	The BioASQ embeddings is created as part of the European project BioASQ. It applies the word2vec tool to abstracts of PubMed (pre-processed). It is one of the in-domain neural embeddings that is publicly available.
PubMed-w2v embeddings [Pyysalo et al., 2013a]	Similar to the BioASQ embeddings, the PubMed-w2v embeddings is created using the Skip-gram model from the word2vec tool. In addition to the abstracts of PubMed, the creators also induce embeddings from PubMed full-text archive (i.e. PMC), as well as the Wikipedia text. Consequently, they release three sets of neural embeddings: PubMed-w2v, PMC-w2v and PubMed-PMC-Wiki-w2v.

Table 2.1 Examples of word representations used in biomedical NLP

2.2.2 Evaluation resources in the biomedical domain

Intrinsic evaluation

In biomedicine, the most commonly used intrinsic evaluation datasets are **MayoSRS** [Pakhomov et al., 2011] and **UMNSRS** [Pakhomov et al., 2011]. MayoSRS consists of 101 clinical term pairs, which are generated manually by a physician. The relatedness of each word pair is rated by nine medical coders and three physicians, based on a ten-point scale (1: closely related, 10: unrelated). On the other hand, UMNSRS consists of 566 and 587 medical word pairs for measuring similarity (UMNSRS-Sim) and relatedness (UMNSRS-Rel) correspondingly. Word pairs included in the dataset are sourced by first selecting all concepts from the Unified Medical Language System (UMLS) with one of three semantic types: *disorders*, *symptoms* and *drugs*, followed by manual filtering from a physician. The degree of association of each data set is then rated by four medical residents from the University of Minnesota Medical School.

Regarding dataset size and coverage, MayoSRS is smaller and emphasizes clinical concepts whereas UMNSRS covers more concepts from different areas of biomedicine (e.g. *drugs* and *disorders*). Both datasets consist of multi-token terms (e.g. '*difficult walking*' and '*aloe vera*'). Nevertheless, both datasets evaluate only noun representations, and there is a lack of evaluation benchmarks for verbs, though they are essential when interpreting the relations between entities mentioned in the biomedical text. Besides, UMNSRS considers both semantic similarity and relatedness whereas MayoSRS only considers the latter. Hence, there are cases where related but semantically dissimilar word pairs (e.g. *pneumonia* and *infiltrate*) are rated higher than those that are both related and similar (e.g. *dyspnea* and *tachypnea*). Consequently, evaluation of representation models on these datasets penalizes the models which capture the fact that *pneumonia* and *infiltrate* are dissimilar. The two issues highlight the importance of developing new intrinsic evaluation resources for biomedicine.

Extrinsic evaluation

Extrinsic evaluation is fundamental because it measures how useful the features in word representations are for downstream applications. Here, we briefly summarize several relevant downstream applications that commonly utilize representation models in biomedical-NLP.

Named entity recognition (NER): It is a subtask of information extraction. It involves extracting and classifying words and phrases in unstructured texts into pre-defined categories such as person names, organizations and locations. For biomedical-NER, this often includes categories such as genes, proteins, chemicals, cell types, diseases, and drugs. NER serves as

Resources	Descriptions
Acromine [Okazaki and Ananiadou, 2006]	An abbreviation dictionary extracted from MEDLINE.
AnatEM [Pyysalo and Ananiadou, 2013]	A corpus annotated with entities from anatomy.
BC2GM [Smith et al., 2008]	A corpus annotated with Gene Mention as released in the Bio-Creative II task.
BC4CHEMD [Krallinger et al., 2015]	A corpus annotated with chemicals and drugs as released in the Bio-Creative IV task.
GENETAG [Tanabe et al., 2005]	A corpus consists of more than 20.000 MEDLINE sentences relating to gene/protein term identification.
GENIA [Kim et al., 2003]	A corpus annotated with terms from the GENIA ontology covering species, chemicals, cell types, genes/protein.
JNLPBA [Kim et al., 2004]	A corpus annotated with technical terms in molecular biology.

Table 2.2 Examples of NER corpora commonly used in the biomedical NLP community

an essential procedure in many biomedical information extraction pipelines. Some of the well-established biomedical-NER corpora are described in Table 2.2.

NER for the biomedical text differs from the general-domain text in several aspects:

1. In biomedical-NER, it is common to have alternate spellings and/or abbreviations for identical entities [Goulart et al., 2011].

2. Biomedical named entities are often composed of long sequences of tokens, making it harder to detect the boundaries for segmentation [Leser and Hakenberg, 2005].

3. Many terminologies are rarely found in general English dictionaries [Krauthammer and Nenadic, 2004].

Representation models have been widely-used in biomedical-NER to provide word semantic features. For examples, Lin et al. [2010] showed that information on word relatedness as provided by representation models could help to determine the correct treatment of acronyms. Moreover, Habibi et al. [2017] showed that using the pre-trained word embeddings on a generic neural-NER tool yielded improvements, achieving state-of-the-art results on several gene and chemical recognition tasks. Similar improvements have also been reported in other biomedical-NER studies using pre-trained word embeddings [Crichton et al., 2017; Tang

Resources	Descriptions
BioLexicon [Thompson et al., 2011]	BioLexcion is a corpus-driven lexicon which contains syntactic and semantic information for nouns and verbs, including the term variants and relations between entities.
MeSH [Lipscomb, 2000]	MeSH is a categorized vocabulary for indexing scientific articles. Maintained by the US NLM.
NCBI-G [Maglott et al., 2005]	NCBI gene is a database of nomenclatures, reference sequences, phenotypes for species.
PASBio [Wattarujeekrit et al., 2004]	PASBio is a lexicon containing predicate-argument structures of 30 verbs as extracted from biological and biomedical lituratures.
UMLS-M [Bodenreider, 2004]	Unified Medical Language System Metathesaurus is a taxonomy database that contains information about biomedical and health related concepts.
UMLS-S [Browne et al., 2000]	Unified Medical Language System SPECIALIST lexicon is a vocabulary database that contains syntactic, morphological, and orthographic information of words commonly found in English and biomedicine.
UniProt [Consortium, 2014]	UniProt is a database of protein sequences and biological functional information.

Table 2.3 Examples of lexical resources commonly used in the biomedical NLP community

et al., 2014]. In this thesis, we will use NER as one of the extrinsic evaluation to measure the quality of our word embeddings.

Automatic lexical acquisition: The biomedical NLP community often makes use of many ontologies, vocabularies, taxonomies and lexical resources. Some examples are provided in Table 2.3.

Building lexicons manually is labour-intensive and time-consuming. There is a need for utilizing NLP techniques to automate this process. Automatic lexical acquisition refers to the automatic or semi-automatic process of building lexicons from unstructured texts. To extract representative features from corpora for lexical acquisition, it typically involves extensive feature engineering and mature NLP pipelines (e.g. POS taggers). Recently, unsupervised methods for inducing distributed word representations or word embeddings have been successfully applied to many NLP tasks. These methods provide an effective way to learn word features automatically from large corpora, easing the need for feature

engineering. In this thesis, we evaluate the utility of these unsupervisedly-learned features in automatic lexical acquisition.

2.3 Lexical resources for NLP

Lexical resources play an important role in NLP because they contain an exhaustive amount of semantic and syntactic properties of words which facilitate accurate information extraction from texts. Several general-purpose lexicons, such as WordNet [Miller, 1995] and FrameNet [Baker et al., 1998], have been developed, providing information about different linguistic properties and relations between words. WordNet groups words into *synsets* (i.e. synonym sets), and documents the semantic relations between synsets, whereas FrameNet groups words by *semantic frames* (i.e. conceptual categories regarding a specific type of event along with its participants) and their predicate-argument patterns. While these resources cover lexical units of various word-types (e.g. nouns, verbs or adjectives), some resources are specifically developed for a particular word-type.

VerbNet [Kipper-Schuler, 2005] is the most extensive verb lexicon in the general domain. It provides class-level information about the semantics and syntax of verbs. The current version of VerbNet (v3.3) consists of 9,344 verbs organized in 329 classes. Each verb class has a detailed description of its syntactic and semantic properties, including the typical types, numbers, and roles of arguments for its member verbs. For example, the members in the *Remove* class (e.g. *delete* and *discharge*) take similar arguments (agent[3]) and can be used to describe similar events. Although VerbNet has a wide-coverage for general-domain NLP applications, it is not designed for specialized domains, such as biomedicine, where verbs tend to have a very different meaning and behaviour than in general English [Ananiadou and Mcnaught, 2006; Venturi et al., 2009]. There are ranges of in-domain terminologies which are rare in general English (e.g. *depolymerize*). Furthermore, the same word in general English and biomedicine can have distinct syntactic and semantic properties. For example, the verb *fire* has different types of arguments and meanings in each domain (*fire a gun* v.s. *fire a neuron*). Hence, there is a need to develop domain-specific resources to support biomedical NLP.

Biomedicine is full of large-scale resources for noun concepts (e.g. entities), including the UMLS Metathesaurus. However, verbs have been neglected, although they are essential for the interpretation of biomedical language. For example, *trigger*, *phosphorylate* and *interact* can be commonly found in documents related to protein-protein reactions. However, they imply different actions between proteins (casual and non-casual). Many biomedical

[3]In VerbNet, agent is defined as argument which intentionally carries out the event

NLP tasks, including relation extraction [Nguyen et al., 2015], use the syntactic structure of verbs (e.g. the predicate-argument structure) to identify relations in biomedical texts. To help biomedical NLP researchers in identifying the particular topic of text and type of relation between entities described in text, it is vital to have resources that contain rich syntactic and semantic information of individual verbs. Nonetheless, the existing lexicons which cover biomedical verbs are usually small in scale and limited to certain sub-domains in biomedicine. For examples, PASBio provides the predicate-argument structures of 30 verbs commonly used in molecular biology, as extracted from over 14,000 MEDLINE abstracts. Additionally, the BioLexicon – a corpus-driven lexicon which contains syntactic and semantic information for verbs – is extracted from the Escherichia Coli (E.Coli) domain, which limits its usefulness to applications that deal with other sub-domains of biomedicine. Alternatively, the UMLS SPECIALIST lexicon is a resource which consists of both general English and medical and health-related vocabularies. Although it provides the typical syntactic patterns for some verbs, there is no statistical information on which patterns are mostly used. Furthermore, it is manually-created and maintained by domain experts, making it difficult to be expanded and extended to other sub-domains in biomedicine. In the next part, we will describe some approaches for constructing verb lexicons, both manually and automatically.

2.3.1 Manual verb classification

Verbs can be grouped based on their shared syntax and semantics. It has been shown that verbs sharing similar meanings tend to also share similar (morpho-)syntactic patterns and thus can be grouped into lexical classes according to a broader range of linguistic properties [Jackendoff, 1992; Levin, 1993; Pinker, 1989]. Such classes can provide a generalized form of representation about groups of verbs sharing similar properties. Representing each verb by its lexical class helps to map many verbs to the same point, and hence reduces the number of parameters to represent them individually. For example, if a corpus has 100,000 verbs, a naive bag-of-words model would require 100,000 parameters to represent all verbs. Instead, if we map them into 100 classes, it only takes 100 parameters (i.e. their class IDs) to represent all.

In literature, the largest and the most widely-used English verb classification is Levin's Verb Classes [Levin, 1993]. In Levin's classification, verbs are primarily grouped in terms of *Diathesis Alternations*, which characterize the number and type of arguments a verb can take. To illustrate, consider two examples for *Material/Product Alternation*:

- That acorn will grow into an oak tree.
- An oak tree will be developed from that acorn.

Here, *Material/Product Alternation* takes two arguments: one raw material type (e.g. acorn) and one product type (e.g. oak tree) respectively. Verbs that share the same or a similar set of Diathesis Alternations, including *develop* and *grow*, are deemed to share certain meaning components and are organized into a semantically coherent class (i.e. *Grow-26.2*). In total, Levin manually analyzed 3,104 verbs and suggested a list of relevant alternations and linguistic features (e.g. morphology) for identifying verb classes.

2.3.2 Automatic verb classification

While manual classifications of a large number of verbs is time-consuming, previous studies have shown that it is possible to automatically acquire verb classes from both general and biomedical texts. They extensively explored and compared a range of syntactic and semantic features useful for verb classification. For examples, Joanis et al. [2008] used verb features such as the frequencies of passive voice usage and the tenses of verbs to classify 845 English verbs into 14 lexical classes, and Li and Brew [2008] used verb features, including the dependency relations between the arguments and the prepositions, to classify 1,300 verbs into 48 Levin's classes. Moreover, Sun [2013] used rich features based on the predicate-argument structure (e.g. verb subcategorization frames and selectional preferences) to classify 192 biomedical verbs into 50 classes. In addition to feature selection, a range of machine learning models have also been considered for verb clustering in the literature, including the distributional kernel method [Ó Séaghdha and Copestake, 2008] and Support Vector Machines [Sun et al., 2008a].

The classification approaches mentioned above are mostly supervised. These approaches assign verbs into one of several pre-defined lexical classes. In contrast, the unsupervised approaches use clustering techniques to induce classes based on the similarity between verbs. For example, Kawahara et al. [2014a,b] used an unsupervised method called the Chinese Restaurant Process [Aldous, 1985] for inducing 699 verb classes from 1,667 verbs. They deployed a two-step clustering model: the candidate verbs were first clustered into groups based on their shared predicate-argument structure, then verbs classes were induced by clustering these features. They suggested that a two-steps clustering method was essential in tackling verb polysemy, producing polysemy-aware verb classes.

Supervised and unsupervised approaches can serve different purposes: an unsupervised approach requires less prior knowledge and can be used to discover new classes in scenarios where no manually-created classification (i.e. training data) is available; however, the resulting classification unavoidably contains noise. In contrast, when relevant training data is available, supervised approaches have an immediate advantage in terms of the precision of verbs they classified, as reflected in previous studies. For example, Sun et al. [2008a]

classified 204 verbs into 17 Levin's classes, using three supervised classifiers (Support Vector Machines, Maximum Entropy and Gaussian method) and one unsupervised method (Pairwise Clustering). They reported a better result when using the supervised method (Gaussian) and a markedly worse result when using the unsupervised method (Pairwise Clustering). Hence, the supervised approach can be useful for supplementing existing classification with additional (and more accurate) members when training data is available.

Sometimes, a small amount of supervision, in the form of labels on the data (seeds), constraints or user feedback, is provided with unsupervised clustering algorithms. This type of approach, commonly known as semi-supervised clustering, not only groups candidates using the classes learned from the seed data, but also extends and modifies the existing set of classes as needed to reflect other regularities in the data. For examples, Vlachos et al. [2009] used the Dirichlet process mixture model to cluster 204 verbs into 17 classes using features like semantic frame distributions and prepositions of verbs. Additionally, they showed that the clustering performance could be further improved when a small number of pairwise constraints indicating if two verbs must link or must not link were added to the algorithm. In addition, Peterson et al. [2016] expanded the work in Kawahara et al. [2014b] by incorporating the annotated VerbNet data to guide the clustering process to predict a VerbNet class for each sense of a verb, which produced a higher-quality clustering.

From the cognitive science perspective, Barak et al. [2014] applied a two-stage Bayesian model to cluster verbs (first based on syntax then on semantic classes) in order to analyze how computational clustering was similar to human verb knowledge generalization. Furthermore, methods which induce verb classes of other languages (e.g. Estonian [Särg, 2017], French [Sun et al., 2008b], Brazilian [Scarton et al., 2014] and German [Roberts and Egg, 2014]) as well as of a particular type of verb (e.g. propositional attitude verbs such as *think* and *want*) have also emerged recently [White et al., 2014].

A wide spectrum of approaches, including supervised, unsupervised and semi-supervised, have been suggested for verb classification. They have been used to classify verb features extracted from corpus data (raw or POS-tagged). Nonetheless, these approaches rely heavily on feature engineering and mature NLP pipelines; these approaches are time-consuming and expensive because expert knowledge is needed to come up with representative features, and therefore do not provide an optimal solution for classification of verbs in specific domains. Work which performs verb classification on automatically-learned features (through neural networks) are emerging recently. For example, Vulić et al. [2017b] performed verb classification across multiple languages based on automatically-learned features. These sets of features were induced unsupervisedly from corpora (without expert knowledge or feature engineering) using neural embeddings. They reported state-of-the-art results in verb

classification across six languages as compared with previous studies that extract features using complicated language-specific resources.

In general, VerbNet classes have supported many NLP tasks, including word sense disambiguation [Brown et al., 2011], information extraction [Schmitz et al., 2012] and text mining applications [Lippincott et al., 2013; Rimell et al., 2013]. It is foreseeable that biomedical NLP can benefit from similar resources. Nevertheless, general lexical resources are not well-suited for biomedical-NLP usage because they provide limited coverage of in-domain terminologies, yet manually developed in-domain lexical resources cover inadequate verbs, and are costly to extend. Thus, it is essential to automate the construction of verb resources in biomedicine. Regarding this, existing approaches rely heavily on feature engineering to extract verb features from the text, which is time-consuming and requires expert knowledge. With the advancement of neural embeddings, we explore how these sets of automatically-learned features can be used to support automatic lexical acquisition.

2.4 Chapter summary

In this chapter, we have briefly described some representation learning approaches, from vector-space models to neural embeddings, which have been commonly used in the NLP community. While neural representations have shown to be useful in supporting many NLP tasks, most studies only involve general-domain texts and evaluation datasets and these results do not necessarily apply to biomedical-NLP tasks. For optimal performance, representation models need to be trained and evaluated with biomedical data.

We have also reviewed the evaluation benchmarks used for measuring the intrinsic and extrinsic properties of word representations. Currently, there is a lack of evaluation datasets in biomedicine that can measure the intrinsic quality of verb representations, though verbs are essential in the meaning interpretation of biomedical sentences.

Verb lexicons, which provide class-level information about the semantics and syntax of verbs, can be a valuable resource for supporting NLP tasks. However, general lexical resources such as VerbNet only provide limited coverage of biomedical verbs, and manually constructed in-domain lexicons (e.g. the UMLS SPECIALIST lexicon) cover a limited number of verbs and are costly to extend. There is a need for utilizing NLP techniques to automate lexical acquisition. Some approaches have been proposed, yet they rely heavily on feature engineering to extract word features from corpora, which is time-consuming and requires expert knowledge. Recently, unsupervised methods for inducing feature representations have been successfully applied to many NLP tasks. In this work, we explore how these models can be used to automate lexical acquisition.

Chapter 3

How to train good word embeddings for biomedical NLP

3.1 Introduction

Offering valuable input to many current NLP applications, word representations have recently been the subject of much research. The current main approach is to embed words into a low-dimensional space using neural networks [Bengio et al., 2003; Collobert and Weston, 2008; Mikolov et al., 2013b; Pennington et al., 2014; Turian et al., 2010]. Such embeddings have offered useful features for many Natural Language Processing (NLP) applications [Bansal et al., 2014; Guo et al., 2014].

Although word embeddings have been studied extensively [Lapesa and Evert, 2014; Lazaridou et al., 2013], most studies only involve general-domain texts and evaluation datasets, and their results do not necessarily apply to domain-NLP (e.g. biomedicine). Conversely, NLP work conducted on specific domains has revealed that representation models tend to be domain-dependent, implying that their training settings (e.g. the scope of the corpora and hyper-parameters) need to be carefully chosen with consideration of the linguistics properties relevant to the domain they aim to support in order to get the maximal performance [Stenetorp et al., 2012].

In this chapter, we conduct large-scale experiments to investigate the optimal training settings for representation learning when applied to biomedical texts. We focus on three critical parameters in the training process: the input corpora, model architectures and hyper-parameter settings. Using the state-of-the-art neural embedding tool (**word2vec**) and both intrinsic and extrinsic evaluations, we present a comprehensive study on how the performance of embeddings changes according to these features.

3.2 Related work

Among different word embedding methods, the Skip-gram (SG) and Continuous Bag-of-Words (CBOW) of Mikolov et al. [2013a] in word2vec tool have consistently achieved cutting-edge results in many NLP tasks, including sentence completion, analogy and sentiment analysis [Fernández et al., 2014; Mikolov et al., 2013a,b]. In the biomedical domain, Muneeb et al. [2015] compared two state-of-the-art word embedding tools: word2vec and Global Vectors (GloVe) on a word similarity task. They found that Skip-gram notably outperforms other models and that its performance can be further improved by using higher dimensional vectors. These two models have also been shown to produce highly competitive representation models in many intrinsic and extrinsic evaluations [Baker et al., 2016; Pyysalo et al., 2013a; Rei et al., 2016; Tsvetkov et al., 2015], as compared to models such as Random Indexing [Kanerva et al., 2000] and Latent Semantic Analysis [Landauer and Dumais, 1997], among others.

Given that the word2vec has been shown to achieve state-of-the-art performance that can be further improved with parameter tuning, we focus on its performance on biomedical data with different inputs and hyper-parameters. We use all available biomedical scientific literature for learning word embeddings using models implemented in word2vec. For intrinsic evaluation, we use the standard UMNSRS-Rel and UMNSRS-Sim datasets [Pakhomov et al., 2011], which enable us to measure similarity and relatedness separately. For extrinsic evaluation, we apply a neural network-based named entity recognition (NER) model to two standard benchmark NER tasks, JNLPBA [Kim et al., 2004] and the BioCreative II Gene Mention task [Smith et al., 2008]. We will now describe them in detail.

3.3 Materials and methods

3.3.1 Corpora and pre-processing

We use two corpora to create word vectors: the PubMed Central Open Access subset (PMC) and PubMed [1]. PMC is a digital archive of biomedical and life science literature, which contains more than 1 million full-text Open Access articles. The PubMed database has more than 25 million citations that cover the titles and abstracts of biomedical scientific publications. A version of PMC articles is distributed in text format whereas PubMed is distributed in XML [2]. Thus, we use a PubMed text extractor [3] to extract title and abstract

[1] The description of the corpora is provided in Section 2.2.1

[2] PubMed in XML: http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/#Data_Mining

[3] PubMed text extractor: https://github.com/spyysalo/pubmed

texts from the PubMed source XML. Both PubMed and PMC were pre-processed with the Genia Sentence Splitter (GeniaSS) [Sætre et al., 2007], which is optimized for biomedical texts. We further tokenize the sentences with the Treebank Word Tokenizer provided by the NLTK python library [Bird, 2006]. The corpus statistics are shown in Table 3.1.

Corpus	Total tokens
PubMed	2,721,808,542
PMC	7,959,548,841
PubMed + PMC	10,681,357,383

Table 3.1 Corpus statistics

3.3.2 Word vectors

Factors that affect the performance of word representations include the training corpora, the model architectures, and the hyper-parameters. To assess the effect of corpora, we generate three variants of each set of word vectors: one from PubMed, one from PMC, and one from the combination of the two (PMC-PubMed). To study how preprocessing affects word vectors, we create vectors from the original text corpora, lower-cased variants, and variants where sentences are shuffled in random order. We further generate two sets of vectors, one by applying the Skip-gram model and one applying the CBOW model, built with the default hyper-parameter values of word2vec. We first evaluate these vectors to determine the better-performing model architecture. Using the better model, we then build vectors by varying values of one hyper-parameter (Table 3.2) and keeping others as default. We repeat the process for every hyper-parameter under examination. We then report the results of these sets of vectors in our intrinsic and extrinsic evaluations.

Parameters	Values
neg	1 / 2 / 3 / **5** / 8 /10 / 15
samp	0 / 1e-1 / 1e-2 / **1e-3** / 1e-4 1e-5 / 1e-6 / 1e-7 / 1e-8 / 1e-9
min-count	0 / **5** / 10 / 20 / 50 / 100 / 200 400 / 800 / 1000 / 1200 / 2400
alpha	0.0125 / **0.025** / 0.05 / 0.1
dim	25 / 50 / **100** / 200 / 400 / 500 / 800
win	1 / 2 / 4 / **5** / 8 / 16 / 20 / 25 / 30

Table 3.2 Hyper-parameters and tested values. Default values shown in bold.

3.3.3 Hyper-parameters

We test the following key hyper-parameters:

1. **Negative sample size (*neg*)**: The representation of a word is learned by maximizing its predicted probability to co-occur with its context words while minimizing the probability for others. However, the normalization of this probability involves a denominator deriving from co-occurrences between words and all their contexts in the corpus, which is time-consuming to compute. To address this issue, Negative sampling is introduced, which only calculates the probabilities on a set number of other randomly chosen negative words (*neg*). With a higher negative sampling, the model tends to capture better word representations by learning from more negative samples, yet, its training process is more computationally costly [Mikolov et al., 2013c].

2. **Sub-sampling (*samp*)**: Sub-sampling refers to the process of reducing occurrences of frequent words. It selects words appearing with a ratio higher than the threshold *samp*, and ignores each occurrence with a given probability. The process is used to minimize the effect of non-informative frequent words in training. Very frequent words (e.g. *in*) are less informative because they co-occur with most words in the corpus. For example, a model can benefit more from seeing an occurrence of *p16* with *CDKN2* than an instance of the frequent co-occurrence of *p16* with *in*.

3. **Minimum-count (*min-count*)**: The minimum-count defines the minimum number of occurrences required for a word to be included in the word vectors. This parameter allows control over the size of the vocabulary and, consequently, the resulting word embedding matrix.

4. **Learning Rate (*alpha*)**: Neural networks are trained by gradually updating weight vectors along a gradient to minimize an objective function. The learning rate controls the magnitude of these updates.

5. **Vector dimension (*dim*)**: The vector dimension is the size of the learned word vector. While a higher dimension tends to capture better word representations, their training is more computationally costly and produces a larger word embedding matrix.

6. **Context window size (*win*)**: The size of the context window defines the range of words to be included as the context of a target word. For instance, a window size of 5 takes five words before and after a target word as its context for training.

3.3.4 Baseline vectors

As baselines, we include the biomedical domain vectors created by Pyysalo et al. [2013a] and Kosmopoulos et al. [2015]. Their corpus statistics are shown in Table 3.3. All of these vectors are built with the Skip-gram model with the default parameter values (see Table 3.2).

Vector	#Token
PMC-PubMed [Pyysalo et al., 2013a]	5,487,486,225 (total)
PMC [Pyysalo et al., 2013a]	2,591,137,744 (total)
PubMed [Pyysalo et al., 2013a]	2,896,348,481 (total)
PubMed [Kosmopoulos et al., 2015]	1,701,632 (distinct)

Table 3.3 Baseline word vectors

3.3.5 Intrinsic evaluation

A standardized intrinsic measure for word representations in the biomedical domain is the UMNSRS-Sim (**Sim**) and UMNSRS-Rel (**Rel**) dataset [Pakhomov et al., 2011] [4]. They have 566 and 587 word pairs for measuring similarity and relatedness (respectively) whose degree of association was rated by humans. The human evaluation on every word pair is converted to a score to determine its degree of similarity, a higher score implying a more similar pair. The range of the score is on an arbitrary scale. While UMNSRS provides scores for determining the degree of similarity for each word pair, we will measure this by calculating the cosine similarity score for each word pair using the learned word vectors. Afterwards, we assess the two scores using Spearman's correlation coefficient (ρ) which compares the ranking between variables regardless of scale in word similarity task. It is a standard metric for word pair ranking comparisons given the word similarity are all cosine scores and we want to compare their rankings. Also, following the standard evaluation protocol, we exclude words that appear only in the reference but not in our models.

3.3.6 Extrinsic evaluation

Given that the ultimate evaluation for word vectors is their performance in downstream applications, we also assess the quality of the vectors by performing NER using two well-established biomedical reference standards: the BioCreative II Gene Mention task corpus (**BC2**) [Smith et al., 2008] and the JNLPBA corpus (**PBA**) [Kim et al., 2004]. Both of these

[4] The description of the dataset is provided in Section 2.2.2

corpora consist of approximately 20,000 sentences from PubMed abstracts manually annotated for mentions of biomedical entity names. Following the window approach architecture with word-level likelihood proposed by Collobert and Weston [2008], we apply a NER tagger built on a simple feed-forward neural network, with a window of five words, one hidden layer of 300 neurons and a hard sigmoid activation, leading to a Softmax output layer. Our word vectors are used as the embedding layer of the network, with the only other input being a low-dimensional binary vector of word surface features.[5] To emphasize the effect of the input word vectors on performance, we avoid fine-tuning the word vectors during training as well as introducing any external resources such as entity name dictionaries. While this causes the performance of the method to fall notably below the state-of-the-art, we believe this minimal approach is an effective way to focus on the quality of the word vectors as they are created by the tool (word2vec).[6] For parameter selection, we estimate the extrinsic performance of word vectors on the development sets of the two corpora using mention-level F-score. For the final experiment with selected parameters, we apply the test sets and evaluation scripts of the two tasks in accordance with their original evaluation protocols.

3.4 Results

We have created vectors with PubMed, PMC and the combination of the two. For each set of vectors, we experiment with different training settings, including the model architectures and hyper-parameters. In this section, we will report their intrinsic and extrinsic evaluation results individually.

3.4.1 Skip-gram vs. CBOW

Table 3.4 (first 2 rows) shows results comparing the Skip-gram (SG) and CBOW models with default hyper-parameter values in intrinsic (left) and extrinsic (right) evaluation, respectively. In general, the Skip-gram vector shows better results than CBOW in both the word similarity task and in entity mention tagging. In CBOW, the representations of a group of context words are learned through predicting one focus word, with the prediction back-propagated averaged over all context words. By contrast, in Skip-gram, the representation of a focus word is learned by predicting every other context word in the window separately, with the prediction error of each context word back-propagated to the target word. This may allow

[5] For example, whether a word starts or contains a capital letter or number. For detailed reference, we make our implementation openly available.

[6] It is an interesting question for future work whether the findings from our extrinsic evaluation apply also to state-of-the-art taggers.

better vectors to be learned as a focus word is trained over more data, but with less smoothing over contexts [7].

Our result is consistent with that of many previous studies, including that of Muneeb et al. [2015], who compared model architectures on different vector dimensions and reported that Skip-gram outperforms CBOW in biomedical domain tasks.

Regarding the effects of shuffling and lower-casing, in word2vec, the learning rate is decayed as training progresses, text appearing early has a larger effect on the model. Shuffling makes the effect of all text (roughly) equivalent. On the other hand, lower-casing ensures that same word but different cases, such as *protein*, *Protein* and *PROTEIN* are normalized (indexed as one term) for training. From Table 3.4, we see that most vectors benefit from lower-casing and shuffling the corpus sentences. In general, there is about 2 point increase over the generic vectors. Although the shuffled-lower vectors perform better, in the following, we report further results based on the unshuffled-text vector to preserve the comparability of results.

	PMC-PubMed		PMC		PubMed	
Model	Sim	Rel	Sim	Rel	Sim	Rel
SG	0.54	0.488	0.507	0.453	0.446	0.497
CBOW	0.435	0.409	0.348	0.351	0.449	0.446
SG-S	0.555	**0.515**	0.54	0.49	0.551	0.502
SG-L	0.542	0.457	0.502	0.424	0.552	0.47
SG-SL	0.543	0.47	0.52	0.459	**0.56**	0.481
CBOW-S	0.415	0.403	0.434	0.424	0.43	0.414
CBOW-L	0.452	0.404	0.447	0.41	0.461	0.425
CBOW-SL	0.461	0.422	0.45	0.39	0.471	0.426

	PMC-PubMed		PMC		PubMed	
Model	BC2	PBA	BC2	PBA	BC2	PBA
SG	60.86	61.89	59.48	62.11	61	62.52
CBOW	55.11	56.97	54.93	58.1	54.25	58.48
SG-S	59.81	62.13	59.23	62.3	60.75	62.11
SG-L	60.52	62.19	59.93	61.64	60.51	**62.64**
SG-SL	**61.33**	62.58	60.23	62.05	61.11	61.65
CBOW-S	51.84	56.78	54.22	58.02	52.82	57.97
CBOW-L	53.72	57.09	54.57	57.51	52.65	57.41
CBOW-SL	52.89	57.15	52.63	56.80	53.21	58.41

Table 3.4 Intrinsic (left, in ρ) and Extrinsic (right, in F-score) evaluation results for vectors with different pre-processing: Original text, Sentence-shuffled (S), lowercased (L), and both (SL) for Skip-gram (SG) and CBOW. Bold indicates the best score for a dataset (Sim: UMNSRS-Sim, Rel: UMNSRS-Rel, BC2: BC2GM and PBA: JNLPBA)

3.4.2 Hyper-parameters

Next, we evaluate six hyper-parameters in word2vec individually, using both intrinsic and extrinsic evaluations. We found that four out of the six hyper-parameters only improve performance notably in the intrinsic task but not the extrinsic one, while one boosts performance

[7] In CBOW, the training contexts for each word are smoothed by considering only the average of its context word vectors within the context window, while Skip-gram trains on every word-context pair

in both tasks to a great extent. Lastly, one of them shows opposite effects on intrinsic and extrinsic evaluations.

Negative sampling (*neg*)

Intuitively, larger values of the *neg* parameter could be expected to benefit the training process by providing more (negative) examples, but we can only see a benefit in the intrinsic result (Figure 3.1, red lines). Looking into the results in Table 3.6, the performance of word vectors on the intrinsic task (left) generally improves as *neg* increases from 1 to 8, whereas extrinsic task performance (right) remains approximately the same.

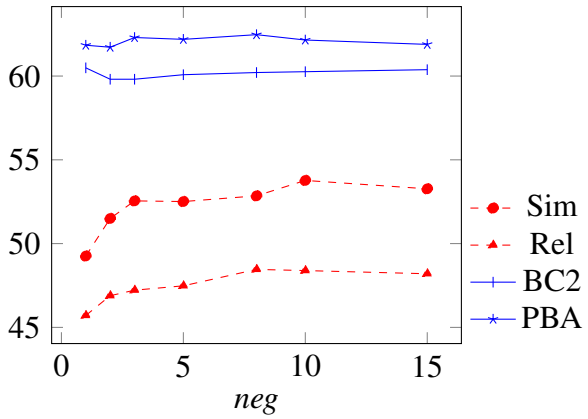


Fig. 3.1 Average intrinsic and extrinsic evaluation results for negative sampling (Unit: ρ: dashed line, F-score: solid line)

	PMC-PubMed		PMC		PubMed	
neg	Sim	Rel	Sim	Rel	Sim	Rel
1	0.52	0.483	0.453	0.405	0.505	0.483
2	0.545	0.493	0.489	0.439	0.511	0.475
3	0.539	0.488	0.506	0.447	0.532	0.482
5	0.538	0.487	0.498	0.444	0.54	0.494
8	0.545	**0.501**	0.497	0.446	0.543	0.507
10	0.543	0.494	0.517	0.459	**0.553**	0.499
15	0.542	0.498	0.514	0.457	0.542	0.491

	PMC-PubMed		PMC		PubMed	
neg	BC2	PBA	BC2	PBA	BC2	PBA
1	60.78	62.29	59.90	61.52	60.80	61.71
2	60.41	62.03	59.44	60.49	59.59	62.63
3	59.37	62.42	59.55	62.02	60.52	62.45
5	60.37	61.90	59.44	62.12	60.44	62.56
8	60.90	62.19	59.49	62.55	60.23	62.68
10	59.65	62.80	59.58	61.61	**61.53**	62.03
15	61.09	61.52	59.92	60.98	60.12	**63.18**

Table 3.6 Detail intrinsic (left, in ρ) and Extrinsic (right, in F-score) evaluation results for vectors with different number of negative samples (default = 5). Bold indicates the best score for a dataset.

Sub-sampling (*samp*)

Regarding sub-sampling, a lower threshold gives more words a probability of being down-sampled. This implies words are less likely to be kept and they will be excluded for training. Intuitively, sub-sampling frequent words not only reduces the computational burden of the training process, but also improves the quality of its resulting word vectors. From Figure 3.2, it appears that sub-sampling has a large effect on the intrinsic task (red lines), where most figures increase substantially before *samp* = 1e-6. After *samp* = 1e-7, figures in both measures drop dramatically. This suggests that the intrinsic tasks are more sensitive to the effect of sub-sampling frequent words than the extrinsic tasks. Some extremely frequent words (e.g. *the*) are effectively non-informative, but other common words may be important for modelling word meaning. Thus, when the sub-sampling threshold decreases continuously, a substantial amount of informative frequent words are downsampled (approximately 10-15% words of the total vocabulary are affected at each sub-sampling points), leading to ineffective learning of the representation.

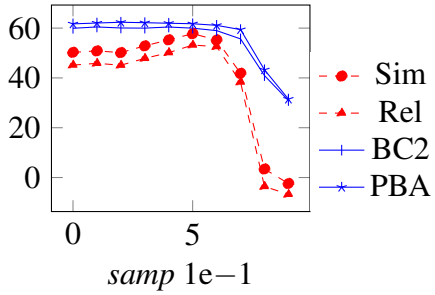


Fig. 3.2 Average intrinsic and extrinsic results for sub-sampling (0 = None)
(Unit: ρ: dashed line, F-score: solid line)

	PMC-PubMed		PMC		PubMed	
samp	Sim	Rel	Sim	Rel	Sim	Rel
None	0.529	0.476	0.465	0.419	0.514	0.451
1e-1	0.542	0.496	0.476	0.42	0.507	0.46
1e-2	0.521	0.464	0.471	0.418	0.513	0.471
1e-3	0.545	0.5	0.497	0.442	0.545	0.494
1e-4	0.56	0.506	0.521	0.459	0.578	0.54
1e-5	0.594	0.542	0.55	0.507	0.589	0.546
1e-6	**0.601**	**0.558**	0.511	0.491	0.546	0.528
1e-7	0.519	0.475	0.401	0.37	0.336	0.306
1e-8	0.09	0.055	0.074	-0.02	-0.06	-0.15
1e-9	-0.07	-0.17	-0.08	-0.18	0.078	0.147

	PMC-PubMed		PMC		PubMed	
samp	BC2	PBA	BC2	PBA	BC2	PBA
None	60.46	61.76	58.83	61.35	60.51	62.00
1e-1	**61.31**	60.99	59.60	62.45	60.47	62.69
1e-2	60.01	62.51	59.86	61.63	60.29	**62.92**
1e-3	60.30	61.99	59.78	61.95	59.87	62.57
1e-4	60.93	62.73	59.87	60.91	60.51	62.22
1e-5	60.58	61.39	60.35	61.26	58.98	62.60
1e-6	60.00	61.67	57.94	60.31	59.02	61.35
1e-7	57.52	61.17	57.04	59.70	52.44	57.34
1e-8	47.35	50.41	44.22	47.23	31.23	32.15
1e-9	33.09	33.13	32.30	32.68	27.40	28.70

Table 3.8 Detail intrinsic (left, in ρ) and extrinsic (right, in F-score) evaluation results for vectors with different sub-sampling (default = 1e-3). Bold indicates the best score for a dataset.

Min-count

Words occurring fewer than *min-count* times will be completely removed from the corpus, resulting in fewer words in the word vectors. From Figure 3.3, most of the results show limited effect for this parameter, excepting a notable increase for PubMed vectors in the intrinsic task (red lines). However, our intrinsic evaluations, following the standard protocol, ignore words that are excluded by *min-count*. Hence, for PubMed vectors, when *min-count* = 400, only about half of the assessment items are used in intrinsic evaluation. This implies that the result in *min-count* > 400 only reflects the representation of frequent words. By contrast, as the out-of-vocabulary rate in extrinsic tasks is about 2.6%, its influence is less notable.

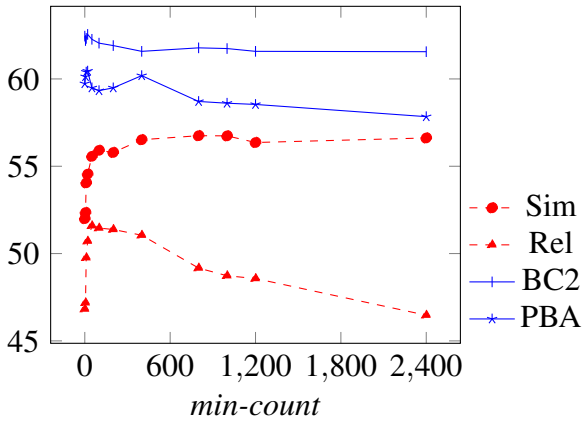


Fig. 3.3 Average intrinsic and extrinsic evaluation results for min-counts (Unit: ρ: dashed line, F-score: solid line)

	PMC-PubMed		PMC		PubMed	
min-count	Sim	Rel	Sim	Rel	Sim	Rel
0	0.543	0.498	0.512	0.444	0.505	0.462
5	0.534	0.485	0.492	0.437	0.544	0.494
10	0.536	0.487	0.528	0.485	0.557	0.521
20	0.531	0.499	0.531	0.492	0.574	0.531
50	0.551	0.523	0.535	0.49	0.581	0.534
100	0.546	0.508	0.553	0.502	0.578	0.534
200	0.547	0.513	0.536	0.49	0.591	**0.538**
400	0.555	0.522	0.543	0.479	0.598	0.531
800	0.55	0.492	0.55	0.467	0.603	0.517
1000	0.551	0.503	0.529	0.443	**0.622**	0.515
1200	0.56	0.506	0.531	0.452	0.601	0.499
2400	0.565	0.485	0.517	0.405	0.616	0.504

	PMC-PubMed		PMC		PubMed	
min-count	BC2	PBA	BC2	PBA	BC2	PBA
0	61.04	62.03	59.73	61.92	59.74	**63.41**
5	60.56	61.83	59.75	61.80	60.52	62.98
10	60.42	62.48	60.22	61.50	60.56	62.98
20	60.64	62.92	60.24	62.17	60.67	62.56
50	**61.32**	62.17	59.58	62.06	59.41	62.59
100	60.59	62.37	58.76	61.47	59.90	62.30
200	59.87	61.39	58.97	61.82	60.00	62.53
400	59.75	62.08	59.95	61.04	60.42	61.62
800	59.35	61.79	59.53	61.75	57.88	61.79
1000	59.98	62.08	58.54	60.98	58.67	62.16
1200	59.26	62.34	58.75	60.74	58.34	61.66
2400	59.49	62.44	58.58	61.54	57.11	60.70

Table 3.10 Detail intrinsic (left, in ρ) and extrinsic (right, in F-score) evaluation results for vectors with different min-count (default = 5). Bold indicates the best score for a dataset.

Learning rate (*alpha*)

The learning process will be unstable if the *learning rate* is too large and will be slow if it is too small. From Figure 3.4, *alpha* = 0.05 appears to be an optimal value, for which most of the vectors have their best or second best results in both evaluations.

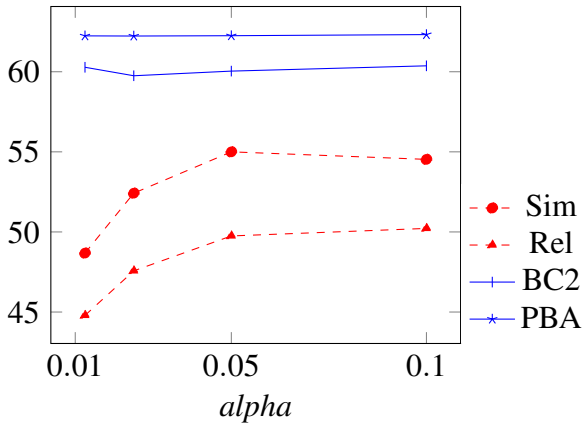


Fig. 3.4 Average intrinsic and extrinsic evaluation results for learning rate
(Unit: ρ: dashed line, F-score: solid line)

	PMC-PubMed		PMC		PubMed	
alpha	Sim	Rel	Sim	Rel	Sim	Rel
0.0125	0.511	0.468	0.442	0.401	0.508	0.475
0.025	0.538	0.492	0.492	0.441	0.543	0.493
0.05	0.55	0.501	0.516	0.46	**0.584**	0.532
0.1	0.542	0.504	0.511	0.46	0.583	**0.543**

	PMC-PubMed		PMC		PubMed	
alpha	BC2	PBA	BC2	PBA	BC2	PBA
0.0125	60.03	61.41	60.24	62.04	60.57	**63.29**
0.025	59.57	61.86	59.86	62.16	59.83	62.68
0.05	59.80	62.86	59.54	61.25	**60.77**	62.65
0.1	60.41	62.38	60.40	61.94	60.30	62.64

Table 3.12 Detail intrinsic (left, in ρ) and extrinsic (right, in F-score) evaluation results for vectors with different learning rate (default = 0.025). Bold indicates the best score for a dataset.

Vector dimension (*dim*)

Intuitively, a higher vector dimension tends to capture better word representation because there is more dimensional space to encode word information. From Figure 3.5, the effect of vector dimension on our vectors is notable in all tasks. We see a large improvement in all evaluations when the vector dimension grows. Although the improvement for intrinsic measures (Figure 3.5, red lines) stops when $dim > 200$, it is evident that an increase from low *dim* gives a very substantial improvement, while the high dimensional representation models appear to have captured additional word properties that are not contributive to our intrinsic and extrinsic tasks.

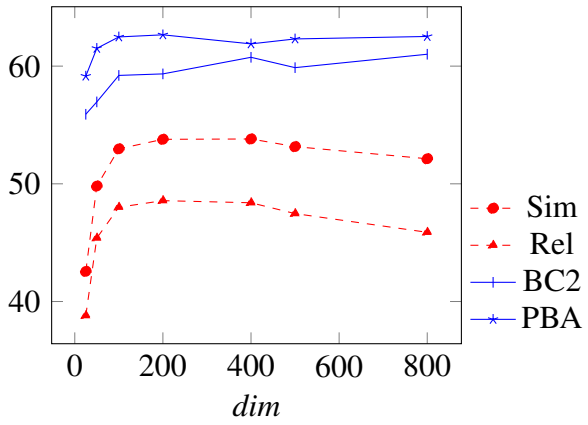


Fig. 3.5 Average intrinsic and extrinsic evaluation results for vector dimension (Unit: ρ: dashed line, F-score: solid line)

	PMC-PubMed		PMC		PubMed	
dim	Sim	Rel	Sim	Rel	Sim	Rel
25	0.426	0.38	0.385	0.346	0.466	0.438
50	0.508	0.461	0.452	0.407	0.534	0.494
100	0.537	0.491	0.509	0.459	0.543	0.491
200	0.552	0.504	0.511	0.459	0.551	0.495
400	**0.562**	0.505	0.518	0.469	0.534	0.477
500	0.553	**0.507**	0.511	0.447	0.531	0.47
800	0.544	0.479	0.51	0.448	0.51	0.45

	PMC-PubMed		PMC		PubMed	
dim	BC2	PBA	BC2	PBA	BC2	PBA
25	56.33	59.14	55.38	58.06	55.77	60.26
50	59.03	61.38	57.24	61.40	57.57	61.75
100	60.81	62.39	60.84	62.17	60.38	62.88
200	61.22	**63.04**	60.13	62.27	**61.24**	62.68
400	61.17	61.57	60.18	61.61	60.54	62.50
500	60.89	62.21	60.81	62.38	61.03	62.36
800	61.00	62.30	60.43	62.34	60.59	62.92

Table 3.14 Detail intrinsic (left, in ρ) and extrinsic (right, in F-score) evaluation results for vectors with different vector dimension (default = 100). Bold indicates the best score for a dataset.

Context window size (*win*)

From Figure 3.6, we observe contradictory results from changing the size of the context window parameter. All three sets of vectors show a notable increase in the intrinsic measures (red lines) when the context window size grows. However, the extrinsic evaluation (blue) shows the opposite pattern: all results in extrinsic tasks have an early performance peak with a narrow window (e.g. *win* = 1), followed by a gradual decrease when window size increases. One possible explanation may be that a larger window emphasizes the learning of domain/topic similarity between words, while a narrow context window leads the representation to primarily capture word function [Turney, 2012]. It is possible that for intrinsic evaluation datasets such as UMNSRS it is more important to model topical rather than functional similarity. Conversely, it is intuitively clear that for tasks such as named entity recognition the modelling of functional similarity such as co-hyponym is centrally important.

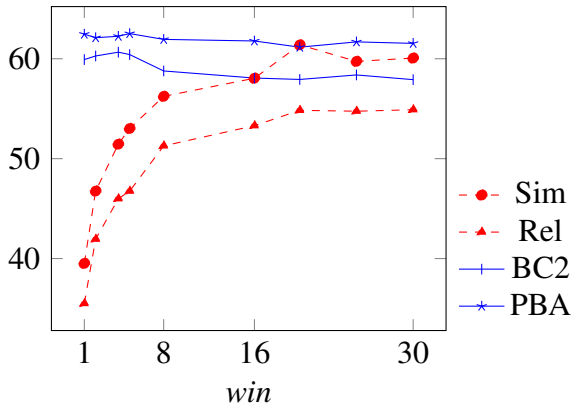


Fig. 3.6 Average intrinsic and extrinsic evaluation results for window size (Unit: ρ: dashed line, F-score: solid line)

	PMC-PubMed		PMC		PubMed	
win	Sim	Rel	Sim	Rel	Sim	Rel
1	0.419	0.377	0.342	0.302	0.425	0.387
2	0.488	0.43	0.422	0.374	0.493	0.454
4	0.528	0.477	0.485	0.425	0.53	0.478
5	0.545	0.494	0.496	0.412	0.55	0.497
8	0.562	0.516	0.544	0.487	0.581	0.536
16	0.589	0.535	0.556	0.506	0.597	0.557
20	**0.66**	0.558	0.562	0.513	0.619	0.574
25	0.6	0.543	0.582	0.531	0.61	0.568
30	0.605	0.541	0.571	0.522	0.627	**0.584**

	PMC-PubMed		PMC		PubMed	
win	BC2	PBA	BC2	PBA	BC2	PBA
1	**61.28**	62.23	60.18	62.44	60.93	62.70
2	60.81	61.74	60.83	61.59	61.11	**63.01**
4	61.29	62.45	60.43	61.43	60.74	62.86
5	59.87	62.25	60.08	62.51	59.47	62.80
8	59.52	61.83	58.78	61.26	60.40	62.74
16	59.82	61.41	59.40	61.30	60.18	62.62
20	59.54	60.80	59.92	60.92	60.02	61.76
25	58.86	60.86	58.91	61.41	58.98	62.79
30	57.83	61.28	57.61	60.53	59.22	62.83

Table 3.16 Detail intrinsic (left, in ρ) and extrinsic (right, in F-score) evaluation results for vectors with context window size (default = 5). Bold indicates the best score for a dataset.

3.4.3 Comparative evaluation

Based on the parameter selection experiments covering three corpora (PMC, PubMed and both), various preprocessing options (normal-text, sentence-shuffled text, lower-cased text), two model architectures (Skip-gram vs CBOW) and six hyper-parameters, we selected the best-performing options for comparative evaluation against the baseline vectors (Table 3.18). Since the size of the context window (*win*) showed contradictory results between the intrinsic and extrinsic tasks, we created vectors for two different values of this parameter. Note that for this comparative evaluation we use the test sets and test evaluation scripts of the two extrinsic tasks, which enables researchers to directly compare our results with the ones reported in literature on these tasks.

Parameters	Values
Corpus	PubMed
Architecture	Skip-gram
neg	10
dim	200
alpha	0.05
samp	1e-4
win	2, 30
min-count	5

Table 3.18 Settings selected for comparative evaluation

Table 3.19 summarizes the results of the comparative evaluation. For our intrinsic tasks, our vectors with *win* = 30 show the best performance, clearly outperforming the baselines as well as our otherwise identically created vectors with *win* = 2. This further supports the suggestion that a higher context window facilitates the learning of domain similarity for the intrinsic task. For extrinsic tasks, while the difference to the baselines is smaller, our vectors with *win* = 2 show the best results for JNLPBA and the second best in BC2GM, while the vectors with *win* = 30 are clearly less competitive.

The comparative evaluation on test set data thus confirms the indications from parameter selection that the context window size has opposite effects on the intrinsic and extrinsic metrics and indicates that our experiments have succeeded in creating two word embeddings of *win* = 2 and *win* = 30 that show promising performance when applied to tasks appropriate for each.

	Sim	Rel	BC2	PBA
PubMed, win 2 (ours)	0.56	0.507	76.89	**64.13**
PubMed, win 30 (ours)	**0.652**	**0.601**	75.51	63.15
Baseline				
Pyysalo et al. (PMC-PubMed)	0.523	0.48	**77.01**	63.6
Pyysalo et al. (PMC)	0.453	0.396	75.48	63.66
Pyysalo et al. (PubMed)	0.549	0.506	76.47	63.66
Kosmopoulos et al. (BioASQ)	0.589	0.509	75.51	62.85

Table 3.19 Intrinsic and extrinsic evaluation with comparison to baseline vectors. Bold indicates the best score for a dataset.

3.5 Discussion

We have created vectors with PubMed, PMC and the combination of the two with a large variety of different models, preprocessing and parameter combinations. In theory, a larger corpus is expected to benefit from the learning of word representations, but we find that in many cases this does not hold, in particular with the combination of PubMed and PMC showing lower results than PubMed alone. We offer two possible explanations for this surprising finding, which contradicts some previous in-domain results. First, we used PMC texts recently introduced by PubMed Central using an incompletely documented extraction process, and preliminary examination suggests that the proportion of non-prose text in this material may be quite high, potentially affecting learning. An alternative explanation may be that the word2vec implementation has a (somewhat hidden) 'reduce-vocab' function that triggers rare-word removal when the size of the corpus crosses certain thresholds: the larger the corpus size, the more aggressive the trimming. Preliminary results suggest that this functionality may have affected PMC-PubMed, our largest corpus, to a larger extent than the other corpora.

3.6 Chapter summary

In this chapter, we investigate how the performance of word vectors changes with different corpora, preprocessing options (normal text, sentence-shuffled text, lower-cased text), model architectures (Skip-gram vs CBOW) and hyper-parameter settings (negative sampling, subsample rate, min-count, learning rate, vector dimension, the context window size). For corpora, sentence-shuffled PubMed texts appear to produce the best performance, exceeding that of the notably larger combination with PMC texts. For hyper-parameter settings, it is evident that performance can be notably improved over the default parameters, but the effects of the different hyper-parameters on performance are mixed and sometimes counterintuitive.

Most importantly, we also observe that changing the sizes of context windows creates contradictory results between intrinsic and extrinsic evaluations. In the next section, we will further investigate if this pattern exists in general-domain text.

Chapter 4

Intrinsic Evaluation of Word Vectors Fails to Predict Extrinsic Performance

4.1 Introduction

With a lack of standardized extrinsic evaluation methods for vector representations of words - word similarity tasks are frequently used as proxies to estimate word quality and intrinsic language properties. Such intrinsic evaluation provides a fast and computationally inexpensive method to measure the quality of representation models.

Word similarity evaluation can measure with human precision how well the notion of semantic similarity is captured in the vector-space representations. It facilitates the estimation of general properties of representation models, which relate to their task performance. Consequently, word similarity evaluation provides a practical means to compare models efficiently before applying them to more elaborate and computationally expensive extrinsic tasks. The underlying assumption is that models that better capture word similarity can, to some degree, perform well on tasks that require a notion of explicit semantic similarity between words like named entity recognition (NER).

Nevertheless, in the previous chapter, we found that most intrinsic datasets in biomedicine are poor predictors of downstream performance. In particular, there is a contradictory estimation between the two sets of evaluations when measuring a series of vector models trained with varying context window sizes. This implies the superior models in intrinsic evaluation may not necessarily perform better on extrinsic tasks (they may even perform worse). Hence, we investigate whether such intrinsic evaluations are suitable for predicting the merits of representations for downstream tasks in general. If not, this observed contradiction could

be a domain-specific effect that possibly only manifests with datasets such as those in the biomedical field.

In this chapter, we study the intrinsic-extrinsic correlation using general-domain datasets. We will focus on vector models of varying context window sizes where the contradictory estimation was found. We base our results on ten word similarity benchmarks and tagger performance on three standard sequence labelling tasks in the general domain, using a variety of word vectors induced from an unannotated corpus of 3.8 billion words from general English.

4.2 Materials and methods

4.2.1 Word vectors

We generated word representations using the **word2vec** implementation of the Skip-gram model [Mikolov et al., 2013a] due to its efficiency when applied to huge corpora. We induced embeddings with varying values of the context window size parameter ranging between 1 and 30 (same as the ones we used in Chapter 3), holding other hyper-parameters to their defaults.[1]

4.2.2 Corpora and pre-processing

We created word vectors by gathering a large corpus of unannotated English text, mainly drawing on publicly available resources identified in the word2vec distribution materials [2].

Table 4.1 lists the text sources and their sizes. We extracted raw text from the Wikipedia dumps using the Wikipedia Extractor [3]; the other sources are textual. We pre-processed all text with the Sentence Splitter and the Treebank Word Tokenizer provided by the NLTK library [Bird, 2006]. In total, there are 3.8 billion tokens (19 million distinct types) in the processed text.

Name	Reference	#Tokens
Wikipedia	Wikipedia [2016]	2,032,091,934
WMT14	Bojar et al. [2014]	731,451,760
1B-word-LM	Chelba et al. [2014]	768,648,884

Table 4.1 Unannotated corpora (sizes before tokenization)

[1] The default parameters are size=100, sample=0.001, negative=5, min-count=5, and alpha=0.025.

[2] `demo-train-big-model-v1.sh`

[3] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

Name	Reference	#Pairs
Wordsim-353	Finkelstein et al. [2001]	353
WS-Rel	Agirre et al. [2009]	252
WS-Sim	Agirre et al. [2009]	203
YP-130	Yang and Powers [2006]	130
MC-30	Miller and Charles [1991]	30
MEN	Bruni et al. [2012]	3,000
MTurk-287	Radinsky et al. [2011]	287
MTurk-771	Halawi et al. [2012]	771
Rare Word	Luong et al. [2013]	2,034
SimLex-999	Hill et al. [2015]	999

Table 4.2 Intrinsic evaluation datasets

4.2.3 Intrinsic evaluation

We performed intrinsic evaluations on the ten benchmark datasets presented in Table 4.2. We followed the standard experimental protocol for word similarity tasks: for each given word pair, we computed the cosine similarity of the word vectors in our representations, and then ranked the word pairs by these values. We finally compared the ranking of the pairs created in this way using the gold standard human ranking Spearman's ρ (rank correlation coefficient).

4.2.4 Extrinsic evaluation

We based our extrinsic evaluation on the seminal work of Collobert and Weston [2008], who explored the use of neural methods for NLP. In brief, we reimplemented the simple *window approach* feedforward neural network architecture proposed by Collobert and Weston [2008], which takes input words in a window of size five, followed by the word embeddings, a single hidden layer of 300 units and a hard tanh activation leading to an output Softmax layer. Besides the index of each word in the embedding, the only other input is a categorical representation of the capitalization pattern of each word.

We trained each model on the training set for 10 epochs using word-level log-likelihood, mini-batches of size 50, and the Adam optimization method with the default parameters suggested by Kingma and Ba [2015]. In order to emphasize the differences between the different representations, we did *not* fine-tune word vectors by back-propagation. In this regard, we diverged from Collobert and Weston [2008] and this led to somewhat reduced performance. We used greedy decoding to predict labels for test data.

To evaluate the word representations in downstream tasks, we used word representations in three standard sequence labeling tasks selected by Collobert et al. [2011]: POS tagging of Wall Street Journal sections of Penn Treebank (**PTB**) [Marcus et al., 1993], chunking of **CoNLL'00** shared task data [Tjong Kim Sang and Buchholz, 2000], and NER of **CoNLL'03** shared task data [Tjong Kim Sang and De Meulder, 2003]. We used the standard train/test splits and evaluation criteria for each dataset. We were evaluating PTB POS tagging using token-level accuracy and CoNLL'00/03 chunking and NER using chunk/entity-level F-scores as implemented in the `conlleval` evaluation script. Table 4.3 shows the basic statistics for each dataset.

Name	#Tokens (Train/Test)
PTB	337,195 / 129,892
CoNLL 2000	211,727 / 47,377
CoNLL 2003	203,621 / 46,435

Table 4.3 Extrinsic evaluation datasets

4.3 Results

Here, we will look at the intrinsic and extrinsic evaluation results of vectors with varying context window sizes. While the different baselines and the small size of some of the datasets makes the intrinsic results challenging to interpret, a clear pattern emerges when holding the result for word vectors of window size 1 as the zero point for each dataset. When examining the average differences, the intrinsic evaluations show higher overall results with increasing window size, while extrinsic performance drops (Figure 4.1).

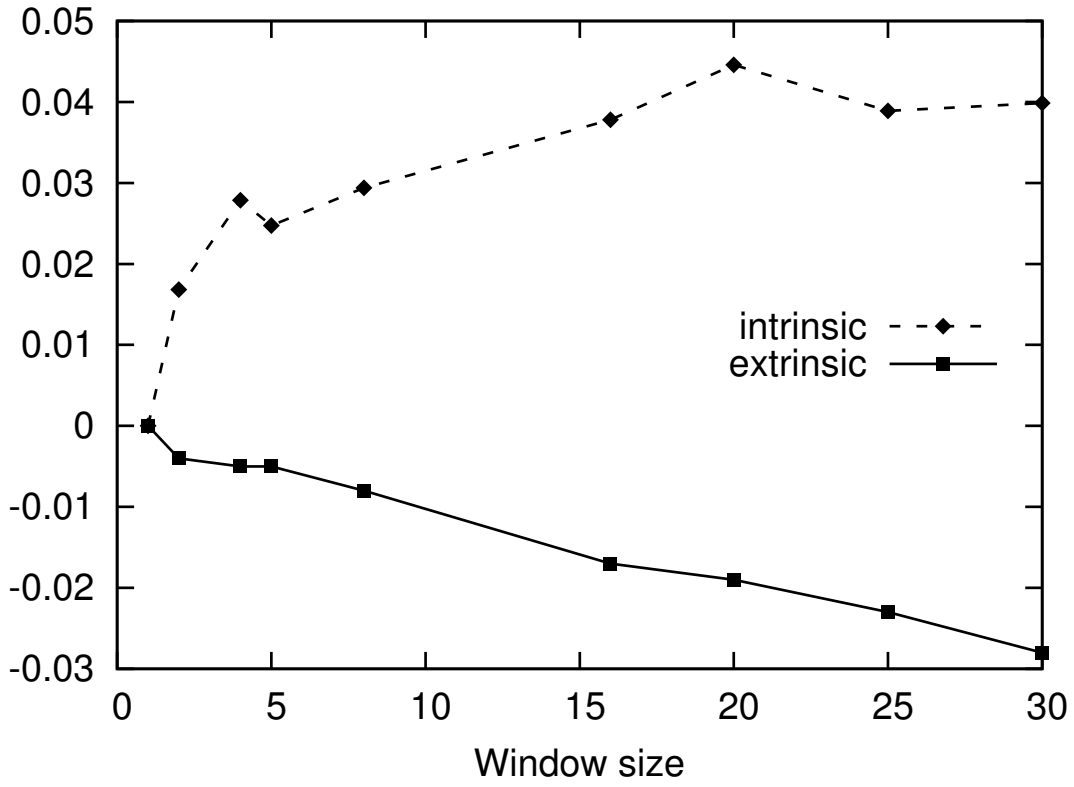


Fig. 4.1 Average difference to performance for window size 1 for intrinsic and extrinsic metrics.

	Window size								
Dataset	1	2	4	5	8	16	20	25	30
WordSim-353	0.6211	0.6524	0.6658	0.6732	0.6839	0.6991	0.6994	**0.7002**	0.6981
MC-30	0.7019	0.7326	0.7903	0.7629	0.7889	0.8114	**0.8323**	0.8003	0.8141
MEN-TR-3K	0.6708	0.6860	0.7010	0.7040	0.7129	0.7222	0.7240	**0.7252**	0.7242
MTurk-287	0.6069	0.6447	0.6403	0.6536	0.6603	0.6580	**0.6625**	0.6513	0.6519
MTurk-771	0.5890	0.6012	**0.6060**	0.6055	0.6047	0.6007	0.5962	0.5931	0.5933
Rare Word	0.3784	0.3893	0.3976	**0.4009**	0.3919	0.3923	0.3938	0.3949	0.3953
YP130	0.3984	0.4089	0.4147	0.3938	0.4025	0.4382	0.4716	0.4754	**0.4819**
SimLex-999	**0.3439**	0.3300	0.3177	0.3144	0.3005	0.2909	0.2873	0.2811	0.2705

Table 4.4 Intrinsic evaluation results (ρ)

	Window size								
Dataset	1	2	4	5	8	16	20	25	30
CoNLL 2000	**0.9143**	0.9070	0.9058	0.9052	0.8982	0.8821	0.8761	0.8694	0.8604
CoNLL 2003	**0.8522**	0.8473	0.8474	0.8475	0.8474	0.8410	0.8432	0.8399	0.8374
PTB POS	**0.9691**	0.9680	0.9672	0.9674	0.9654	0.9614	0.9592	0.9560	0.9531

Table 4.5 Extrinsic evaluation results (F-score for CoNLL datasets, accuracy for PTB)

Tables 4.4 and 4.5 present the results of the intrinsic and extrinsic evaluations respectively. Looking at the individual datasets, the preference for the smallest window size is consistent across all three tagging tasks (Table 4.5), however only one out of the eight intrinsic evaluation datasets, Simlex-999, selects this window size, with the majority clearly favoring larger window sizes (Table 4.4). This contradictory pattern is consistent with the one we observed previously when conducting experiments using biomedical vector models (in Chapter 3). Here, we empirically illustrate that the contradiction is neither a domain-specific issue nor merely a scenario that appeared by coincidence.

	CoNLL 2000	CoNLL 2003	PTB POS
WordSim-353	-0.90	-0.75	-0.88
MC-30	-0.87	-0.77	-0.90
MEN-TR-3K	-0.98	-0.83	-0.97
MTurk-287	-0.57	-0.29	-0.50
MTurk-771	0.28	0.37	0.27
Rare Word	-0.57	-0.29	-0.50
YP130	-0.82	-0.93	-0.50
SimLex-999	**1.00**	**0.85**	**0.98**

Table 4.6 Correlation between intrinsic and extrinsic measures (ρ)

To further quantify this discrepancy, we ranked the word vectors from highest- to lowest-scoring according to each intrinsic and extrinsic measure and evaluated the correlation of each pair of these rankings using ρ. The results are striking (Table 4.6): six out of the eight intrinsic measures have *negative* correlations with all the three extrinsic measures, indicating that when selecting among the word vectors for these downstream tasks, it is *better to make a choice at random* than to base it on the ranking provided by any of the six intrinsic evaluations.

4.4 Discussion

Only two of the intrinsic evaluation datasets showed positive correlation with the extrinsic evaluations: MTurk-287 (ρ 0.27 to 0.37) and SimLex-999 (ρ 0.85 to 1.0). One of the differences between the other datasets and the high-scoring Simlex-999 is that it explicitly differentiates similarity from relatedness and association. For example, in the MEN dataset, the nearly synonymous pair (*stair, staircase*) and the highly associated but non-synonymous pair (*rain, storm*) are both given high ratings. However, as Hill et al. [2015] argues, an evaluation that measures semantic similarity should ideally distinguish these relations and credit a model for differentiating correctly that (*male, man*) are highly synonymous, while (*film, cinema*) are highly associated but dissimilar.

This distinction is known to be relevant to the effect of the window size parameter. A larger window not only reduces sparsity by introducing more contexts for each word, but is also known to affect the trade-off between capturing *domain* similarity vs. *functional* similarity: Turney [2012] notes that with larger context windows, representations tend to capture the *topic* or *domain* or a word, while smaller windows tend to emphasize the learning of word functions. This is because the role/function of a word is categorized by its proximate

	Window size								
Dataset	1	2	4	5	8	16	20	25	30
WS-Rel	0.5430	0.5851	0.6021	0.6112	0.6309	0.6510	0.6551	**0.6568**	0.6514
WS-Sim	0.7465	0.7700	0.7772	0.7807	0.7809	**0.7885**	0.7851	0.7789	0.7776

Table 4.7 Intrinsic evaluation results for WS-Rel and WS-Sim (ρ)

syntactic context, while a large window captures words that are less informative for this categorization. For example, in the sentence *Australian scientist discovers star with telescope*, the context of the word *discovers* in a window of size 1 includes *scientist* and *star*, while a larger context window will include more words related by topic such as *telescope* [Levy and Goldberg, 2014]. The association of large window sizes with greater topicality is also discussed by Hill et al. [2015] and Levy et al. [2015].

This phenomenon provides a possible explanation for the preference for representations created using larger windows exhibited by many of the intrinsic evaluation datasets: as these datasets assign high scores also to word pairs that are highly associated but dissimilar, representations that have similar vectors for all associated words (even if not similar) will score highly when evaluated on the datasets. A large window is beneficial if there is no need for the representation to make the distinction between similarity and relatedness. On the other hand, the best performance in the extrinsic sequence labelling tasks comes from window size 1. This may be explained by the small window facilitating the learning of word function, which is more important for the POS tagging, chunking, and NER tasks than the topic. Similarly, given the emphasis of SimLex-999 on capturing genuine similarity (synonyms), representations that assign similar vectors to words that are related but not similar will score poorly. Thus, we observe a decreasing trend with increasing window size for SimLex-999.

To further assess whether this distinction can explain the results for an intrinsic evaluation dataset for representations using small vs large context windows, we studied the relatedness (WS-Rel) and similarity (WS-Sim) subsets [Agirre et al., 2009] of the popular WordSim-353 reference dataset (included in the primary evaluation). Table 4.7 shows the performance of representations with increasing context window size on these subsets. In general, both show higher ρ with an increasing context window size. However, the performance in the relatedness subset increases from 0.54 to 0.65 whereas that in similarity only increases from 0.74 to 0.77. Thus, although the similarity subset did not select a small window size, the lesser preference for a large window compared to the relatedness subset lends some support to the proposed explanation.

4.5 Chapter summary

One of the primary goals of intrinsic evaluation is to provide insight into the quality of representation before it is used in downstream applications. However, in this chapter, we found that the majority of word similarity datasets in the general domain fail to predict which representations will be successful in sequence labelling tasks, with only one intrinsic measure, SimLex-999, showing high correlation with extrinsic measures. In concurrent work (as described in Chapter 3), we have also observed a similar effect on biomedical domain tasks and word vectors. We further considered the differentiation between relatedness (association) and similarity (synonymy) as an explanatory factor, noting that the majority of intrinsic evaluation datasets (in both general and biomedical domains) do not systematically make this distinction. Our results underline once more the importance of such distinction, as well as including also extrinsic evaluation when assessing NLP methods and resources.

Chapter 5

Bio-SimVerb and Bio-SimLex: wide-coverage evaluation sets of word similarity in biomedicine

5.1 Introduction

With the growing use of word representations in Natural Language Processing (NLP) tasks, the quality and consistency of their evaluations have become pivotal in their development [Faruqui et al., 2016; Tsvetkov et al., 2015]. Existing evaluation protocols can be broadly categorized into two groups: intrinsic and extrinsic.

While several intrinsic evaluation resources have recently been developed for the general domain, similar resources for biomedicine currently suffer from notable shortcomings. First, they fail to distinguish between the concepts of semantic similarity (e.g. *dyspnea* and *tachypnea*) versus semantic relatedness (e.g. *pneumonia* and *infiltrate*). In the previous chapter, we showed that such distinctions are important predictors concerning the usefulness of representation models in extrinsic tasks such as named entity recognition (NER) and part-of-speech tagging. Consequently, evaluation datasets (e.g. SimLex-999, [Hill et al., 2015]) which make such distinctions in their design protocols proved to be a better predictor of vector performance in extrinsic tasks. Second, there is currently a lack of evaluation datasets for semantic representation of biomedical verbs, despite their indispensable role in the interpretation of biomedical language.

To address these two shortcomings, in this chapter, we describe the creation of two new resources for evaluation of word representations in biomedicine: **Bio-SimVerb** and **Bio-**

SimLex. These are wide-coverage and easy-to-implement evaluation resources for analyzing verb and noun representations, respectively.

5.2 Dataset design

This section describes the design protocols of Bio-SimVerb and Bio-SimLex.

5.2.1 Choice of words

Samples/words in Bio-SimVerb (verbs) and Bio-SimLex (nouns) are collected from a pre-processed PubMed Central Open Access subset (PMC), which is distributed by Hakala et al. [2016]. POS tags and tokens in this resource are generated using the BLLIP constituency parser [Charniak and Johnson, 2005], trained on a biomedical corpus [Mcclosky, 2010]. The resource covers over 1.4M full articles with more than 388M parsed sentences.

After retrieving all samples from the PMC, we remove all multi-word expressions (e.g. '37 degrees C') and auxiliary verbs (e.g. 'must'). We also filter out noise, such as symbols (e.g. '$<$'), numbers (e.g. '2010'), strings too short to be reliably understood (e.g. 'a', 'v', 'b1') and Greek letters ('α'). In the next step, we use the Bio-lemmatizer [Liu et al., 2012] for lemmatization of non-lemmas (e.g. 'gone', 'went', 'cells'). We also normalize words with the British English spelling into their American English variants for consistency. We exclude terms occurring less than five times, as they are most likely uninformative. These steps filter down our samples from 20,281 to 6,425 verbs, and from 1,339,806 to 217,425 nouns. We have then invited two researchers working in biomedical NLP to determine whether these terms are mostly used in the biomedical or general domains. We exclude samples with ambiguous and multiple usages in both domains (e.g. 'play', 'fire'). Consequently, 526 and 483 verbs, plus 1,312 and 840 nouns, are categorized as commonly used in the biomedical domain and general domain, respectively. Several example words from both domains are provided in Table 5.1.

Biomedical	General
depolymerize	automate
electrophorese	study
phosphorylate	argue
centrosome	idea
pathophysiology	people
endothelium	river

Table 5.1 Biomedical- and general-domain word samples in Bio-SimVerb and Bio-SimLex.

To show that the selected biomedical terms are domain-specific, we have examined individual samples based on their frequency differences in the biomedical and general English texts. We compare the relative frequency of our samples in PMC with that in the British National Corpus (BNC) [Consortium, 2007]. We calculate the Spearman's correlation (ρ) between their frequency ranking in these corpora. The result is only a weak correlation: $\rho = 0.39$, implying that the usage patterns of words in these areas are distinct.

To ensure broad coverage of samples from various areas of biomedicine, we keep track of every journal where a sample appears. These journals are categorized by 125 Broad Subject Terms [NLM, 2017], which are assigned by the U.S National Library of Medicine (NLM) to MEDLINE journals to describe the journal's overall scope and nature. For each sample obtained from PMC, we record the PMCIDs of all the journals in which it appears. We then map the PMCIDs to their corresponding Broad Subject Terms. Consequently, we generate the distribution of Broad Subject Terms for individual samples based on their occurrence in journals. Since one sample can appear in journals with different Broad Subject Terms, we assign the one with the highest occurrence frequency.

The use of Broad Subjects Terms and the examination of frequency for our samples demonstrate the extensive coverage of words in Bio-SimLex and Bio-SimVerb originating from different biomedical areas.

5.2.2 Constructing concept pairs

Next, we sketch the process of constructing concept word pairs for the final annotation. In general, our dataset is made up of *quarters* of word pairs: around 250 associated pairs and 250 unassociated pairs are from the biomedical domain; 250 associated pairs and 250 unassociated pairs are from the general domain.

Concept pairs from the biomedical domain

To form associated pairs in the biomedical quarter, we use two publicly available semantic resources:

1. **SPECIALIST Lexicon** As part of the Unified Medical Language System (UMLS), the SPECIALIST Lexicon provides information about common English vocabulary and biomedical terms found in MEDLINE as well as in the UMLS Metathesaurus. Each entry in SPECIALIST includes syntactic (e.g. *form* and *forms*), morphological (e.g. *localised* and *localized*), and semantic variants (e.g. *breathe* and *respire*). To form associated pairs, we pair up our concepts randomly sampled from the PMC. From these random pairings, we have detected that 121 nouns and 80 verb synonymous

Ontology	Reference
Chemical Entities of Biological Interest (ChEBI)	Hastings et al. [2013]
Gene Ontology (GO)	Ashburner et al. [2000]
NCI Thesaurus (NCIT)	Golbeck et al. [2011]
Foundational Model of Anatomy (FMA)	Rosse and Mejino Jr [2008]
Disease Ontology (DOID)	Kibbe et al. [2014]
Uberon multi-species anatomy ontology (UBERON)	Mungall et al. [2012]
Plant Ontology (PO)	Walls et al. [2012]
Plant Phenotypes and Traits (PATO)	Gkoutos et al. [2004]
Ontology for Biomedical Investigations(OBI)	Brinkman et al. [2010]
Molecular Process Ontology (MOP)	Batchelor [2017]
Zebrafish anatomy and development (ZFA)	Van Slyke et al. [2014]
Protein modification (PSI-MOD)	Montecchi-Palazzi et al. [2008]
Common Anatomy Reference Ontology (CARO)	Haendel et al. [2008]
Xenopus anatomy and development (XAO)	Segerdell et al. [2008]

Table 5.2 14 Ontologies used for sampling synonymous pairs in Bio-SimVerb and Bio-SimLex

pairs appear in SPECIALIST. These pairs, together with pairs found in other resources (described in the next section), are included in Bio-SimLex and Bio-SimVerb after a manual inspection by our biomedical NLP researchers.

2. **The Open Biomedical Ontologies** The Open Biomedical Ontologies Foundry [Smith et al., 2007] creates a collection of ontologies for shared use across different biological and medical domains. Each ontology provides a fine-grained representation of similar entities within a sub-domain. We use synonyms, as well as sibling entities (i.e., entities sharing the same parent node in an ontology), provided in 14 ontologies (see Table 5.2) as the reference for finding synonymous pairs. Since many terms in these ontologies are nominalized forms of verbs (e.g. *phosphorylation* instead of *phosphorylate*), we first include all word forms for every term in the Ontologies by querying its morphological variants in the SPECIALIST Lexicon. Following that, we match our random pairs to the synonymous pairs found in these ontologies.

From our random pairs, we find 506 (nouns) and 287 (verbs) synonymous pairs in these ontologies, together with the semantic pairs previously found in SPECIALIST (nouns: 121 and verbs: 80). This yields a total of 627 noun pairs and 367 verb pairs. They are all inspected by our biomedical NLP researchers manually to ensure that pairs are associated in a biomedical sense. The experts agree that 247 noun pairs and 250 verb pairs have an association: this forms the quarter of associated word pairs in the biomedical domain.

Using a set of random pairs which are not found in any of the two semantic resources, we randomly sample 247 noun pairs and 250 verb pairs. We also consult the experts to ensure that all these pairs are either minimally associated or completely unassociated. They form the quarter of unassociated pairs in the biomedical domain.

Concept pairs from the general domain

Bio-SimLex and Bio-SimVerb contain 494 noun pairs and 500 verb pairs that are commonly used in English. We now describe how to form such word pairs from our samples, with reference to the USF norms dataset [Nelson et al., 2004] containing word association norms.

The USF Norms Dataset The USF dataset is the largest database of free word association collected in word norming experiments for English. It has 72,000 associated word pairs. The pairs are created by presenting one of 5,000 cue concepts to human subjects and then recording their first associated words. This way, each concept is rated by over 10 participants, yielding a set of associates for every concept. The forward and backward association strengths between a concept and its associates are reported in the USF. It includes both related but dissimilar pairs (e.g. *player/team*), as well as similar pairs (e.g. *to wash/to rinse*).

In our case, we again pair up concepts randomly sampled from the PMC. From these pairs, we extract 247 noun pairs and 250 verb pairs represented in the USF: we require the pairs to be assessed by more than 10 USF participants, as well as to have both forward and backward association strengths assigned. These two filtering conditions not only ensure that two words in a pair have a degree of semantic association but also guarantee that the association link is bidirectional. A similar sampling procedure is used in the construction of general-domain benchmarks including SimLex [Hill et al., 2015] and SimVerb [Gerz et al., 2016]. Finally, we also extract 247 noun pairs and 250 verb pairs not present in the USF to form the quarter of unassociated words pairs in the general domain.

5.2.3 Concept pair scoring

Bio-SimLex and Bio-SimVerb consist of 988 noun pairs and 1,000 verb pairs respectively. The similarity between concepts in each pair is determined by twelve annotators who all have a background in biology. Seven annotators are undergraduate or post-graduate students in the Biology School, University of Cambridge, while the remaining five are biologists working at the Institute of Environmental Medicine, Karolinska Institutet [1]. The similarity is assessed

[1]We did not keep track of the English proficiency of the annotators provided that the main concern in this study is their domain expertise regarding the biomedical verb semantic, yet, we will keep a note on this and will address it in future study

on a scale of 0-6, where 0 is assigned to completely unrelated concepts, and 6 represents highly synonymous concepts. The same scale is used in the construction of SimVerb and SimLex.

We adopt the annotation protocol established in prior work on SimVerb and SimLex: the annotators are instructed to assign low scores to related but dissimilar word pairs (e.g. *drug/pharmacy*). In each data set, we randomly select 50 pairs to serve as a consistency set. This set is used to detect possible variation between annotators and data subsets. We then divide all pairs from Bio-SimVerb and Bio-SimLex into two groups, containing approximately 600 pairs each. Out of these 600 pairs, 500 are unique to each group, and 50 pairs are from the consistency set, included in both groups. Another 50 are duplicate pairs displayed to each rater twice to detect his or her inconsistent annotations. Each annotator rates one group. Consequently, each pair is rated by six participants in total. The final survey is implemented so that each rater sees 120 pairs per page on the interface: 100 unique ones, 10 from the consistency set, and 10 duplicate pairs.

The pairs are rated by moving a slider. The participants are explicitly asked to give the same rating to the same pairs for consistency [2]. Furthermore, we also monitor for suspicious rating patterns (e.g., randomly alternating between two ratings). If a participant uses a single rating for ten consecutive questions, we issue a warning to the participant as a reminder to pay attention throughout the survey.

[2] 50 are duplicate pairs displayed to each rater twice to detect his or her inconsistent annotations

5.3 Experimental setup

5.3.1 Word representation models

To evaluate Bio-SimVerb and Bio-SimLex, we apply a range of popular word representation models. All models are trained on a corpus of PubMed abstracts consisting of approximately 2.7 billion tokens (11,980,338 types). The common hyper-parameters shared by these models are standardized to the values shown in Table 5.3, while parameters specific to individual models are kept at their defaults [3].

Parameters	Values
Context window size	5
Vector dimension	200
Learning rate	0.05
Negative sampling	5
Min-count	5
Sampling rate	1e-5

Table 5.3 Hyper-parameter values for word representation models. Parameters specific to individual models are set to their defaults.

[3]Throughout the study in Chapter 3, 4 and 5, we used the same set of default values, as suggested in the word2vec package. For the description of individual hyper-parameters, one can refer to Chapter 3 (Section 3.3.3)

We included five representation models which are commonly used in the field:

1. **Skip-gram (SG) and Continuous Bag of Words (CBOW)** The word2vec tool [Mikolov et al., 2013a] has been shown to produce highly competitive representation models in many intrinsic and extrinsic tasks [Baker et al., 2016; Pyysalo et al., 2013a; Rei et al., 2016; Tsvetkov et al., 2015]. Hence, the representation models used in these experiments are mostly built on the Skip-gram and CBOW architectures. In Skip-gram, the vector for each word is learned by predicting other words within a given context window. Conversely, in the CBOW model, a word is predicted given its context [4].

2. **Structured Skip-gram (SSG)** Based on the SG model, Ling et al. [2015a] proposed an extension, Structured Skip-gram (SSG), which captures word order information. In the SSG model, the vector of each word is learned by predicting not only its context words but also its relative position. This model has shown improvement in various syntactic tasks as compared to original SG models [Ling et al., 2015a].

3. **CBOW with attention (Attention)** Based on the CBOW architecture, Ling et al. [2015b] introduced an attention mechanism which finds the contextual words that are most relevant for each prediction. Their results showed that this model could benefit both semantic and syntactic tasks [Ling et al., 2015b].

4. **SG with dependency-parse (Dependency)** Levy and Goldberg [2014] proposed using dependency-parsed texts to help representation learning in word2vec so that learning includes syntactic dependencies and is not restricted to a fixed context window. This model has been shown to better capture the functional similarity of words than the original SG models [Levy and Goldberg, 2014].

In addition to applying the above models, we also include seven previously released word representations in both the general and biomedical domains:

1. **Paragram, Paragram+CF, Symmetric, CBOW-general and Dep-general** Biomedical representation models are domain-specific, which implies that the word semantics they capture can be different from those in the general domain. To study this, we also include five general-domain representation models previously benchmarked on SimVerb and SimLex: a model learned from the paraphrase database (**Paragram**) [Wieting et al., 2015] and its extension fine-tuned by linguistic constraints from other knowledge

[4] The descriptions of SG and CBOW are provided in Section 2.1.1

resources (**Paragram+CF**) [Mrkšic et al., 2016], a model learned from symmetric-patterns in corpus such as 'x *rather than* y' and '*either* x *or* y' (**Symmetric**) [Schwartz et al., 2015] as well as CBOW (**CBOW-general**) and dependency models (**Dep-general**).

2. **PubMed-w2v and BioASQ** created by Pyysalo et al. [2013a] and Kosmopoulos et al. [2015] (resp.) and built with the SG model with vector dimension of 200 and a context window size of 5. [5] They denote the biomedical domain vectors which have been popularly used in literature [Björne, 2014; Björne and Salakoski, 2018].

5.3.2 Intrinsic evaluation

We perform intrinsic evaluations on the **MayoSRS** [Pakhomov et al., 2011] and UMNSRS word similarity datasets [Pakhomov et al., 2010]. For UMNSRS, We use its **UMNSRS-Sim** and **UMNSRS-Rel** subsets as our references. They have 566 and 587 word pairs for measuring similarity and relatedness (respectively) whose degree of association was rated by participants from the University of Minnesota Medical School. We use the standard experimental protocol for word similarity tasks: for each word pair in a dataset, we compute the cosine similarity of the two word representations and rank the word pairs by these values. We then compare the ranking against a ranking based on human similarity scores using Spearman's correlation (ρ).

5.3.3 Extrinsic evaluation

We assess our representation models using a NER task with four established corpora: the Anatomical Entity Mention corpus (**AnatEM**) [Pyysalo and Ananiadou, 2013], the BioCreative II Gene Mention task corpus (**BC2**) [Smith et al., 2008], the BioCreative IV Chemical and Drug NER corpus (**CHEMD**) [Krallinger et al., 2015] and the JNLPBA corpus (**PBA**) [Kim et al., 2004]. The NER model follows the simple window-based feed-forward network architecture proposed by Collobert and Weston [2008]. Table 5.4 shows the hyper-parameters used in this model.

The model input consists of the vectors of words within a context window, connected to a single hidden layer with a hard tanh activation, leading to an output Softmax layer for predicting labels for named entities. Performance is evaluated using entity-level F-score as implemented in the standard `conlleval` evaluation script.

[5] The descriptions of these models are provided in Section 3.3.4

Parameters	Values
Vector dimension	200
Hidden layer dimension	300
Context window size	5
Learning rate	0.01
Dropout probability	0.2
Epochs	20
Minibatch size	50

Table 5.4 Hyper-parameters used in NER

5.4 Results

5.4.1 Inter-rater reliability

In this study, each annotator rated one sub-group of pairs in Bio-SimVerb and Bio-SimLex. We used the previously published implementation from the SimLex and SimVerb studies to estimate inter-annotator agreement (IAA). In this implementation, IAA-1 computes the average pairwise Spearman's correlation (ρ) of ratings for each annotator with the ratings of all the other annotators. To smooth individual rater effects, we also include IAA-2 (mean), which computes the Spearman's correlation of individual annotators' ratings with the average ratings of all the other annotators within the same group.

We first computed IAA-1 between the ratings of all annotators on the consistency set. Based on these results, we removed the annotations of one outlier whose IAA-1 was considerably lower than the average IAA-1 of all the other annotators from the data. After that, we computed IAA-1 and IAA-2 between annotators rating the same group. The average IAA-1 and IAA-2 for Bio-SimVerb are 0.65 and 0.69 respectively, whereas the results for Bio-SimLex are 0.72 (IAA-1) and 0.78 (IAA-2). We then calculated the average of all ratings from the accepted annotators for each pair and scaled the scores linearly from the 0-6 to the 0-10 interval to match other datasets such as MayoSRS. Following the standard protocol, the similarity score for a representation model is computed using cosine similarity for each word pair, and the performance of the model is then measured by the Spearman's correlation between its ranking of the pairs and the human ranking.

Model	UMN-rel	UMN-sim	Mayo	Bio-SimVerb	Bio-SimLex	CHEMD	BC2	AnatEM	PBA
Attention	0.5248	0.5551	**0.6113**	0.471	0.7155	79.11	65.91	80.49	62.3
SSG	0.5189	0.552	0.6003	**0.4744**	0.7181	79.62	67.3	81.3	63.78
SG	**0.5767**	**0.6271**	0.5744	0.4638	0.7151	81.37	70.2	81.32	**65.16**
CBOW	0.5	0.5348	0.5146	0.4367	0.702	78.41	64.05	80.3	61.9
Dependency	0.3934	0.4622	0.3445	0.3978	**0.7436**	**83.69**	**71.43**	**82.4**	65.01
PM-w2v	0.506	0.549	0.5133	0.4376	0.6984	80.71	67.4	81.1	64.86
BioASQ	0.5092	0.5893	0.4729	0.4228	0.6982	56.95	48.86	53.34	50.51

Table 5.5 Intrinsic (left 5 columns, in ρ) and extrinsic scores (right 4 columns, in F-score) of different representation models trained on the biomedical corpus.

5.4.2 Performance of representation models on intrinsic evaluation datasets

Table 5.5 shows the intrinsic (left 5 columns) and extrinsic scores (right 4 columns) of the different representation models. To address ties in human scores in intrinsic evaluations, we use the Scipy implementation (v0.19) [Jones et al., 2001] to compute the tie-corrected Spearman's correlation as suggested by Kendall and George [1955]. This correction handles the ties by averaging the uncorrected correlation values over all possible valid (without ties) rankings of the underlying variable. To account for variance in neural networks due to their random initialization, we run three trials for all extrinsic tasks and report their averages.

In general, scores are higher in Bio-SimLex than in Bio-SimVerb for all representation models, indicating that it is still difficult for current models to capture verb semantics. In particular, the score of the dependency model is low in Bio-SimVerb, which implies that using dependency parses to reach beyond bag-of-word context may not contribute equally to the representation learning of verbs and nouns. To a large extent, to identify learning algorithms that are useful for learning word-type specific representations (e.g. verbs), resources for the evaluation of specific word-types are a necessity.

	CHEMD	BC2	AnatEM	PBA
Bio-SimVerb (ours)	0.2	0.18	0.29	0.24
Bio-SimLex (ours)	**0.53**	**0.6**	**0.46**	**0.48**
Baseline				
UMN-rel	-0.15	-0.14	-0.08	-0.07
UMN-sim	-0.38	-0.34	-0.34	-0.3
Mayo	0.08	0.04	0.18	0.12

Table 5.6 Pearson's correlation between word-similarity/Bio-SimVerb and Bio-SimLex scores and the NER tasks evaluated on biomedical representation models trained with different approaches. None of the scores are statistically significant. (**Bold**: best scores)

5.4.3 Correlation between intrinsic and extrinsic scores

From Table 5.5, we observe that there is variation in the performance of different representation models across various tasks. For example, the best-performing model in MayoSRS is the attention model, whereas the dependency model performs best in most NER tasks. To study if our datasets can predict extrinsic performance, we compute the Pearson's correlation (r) to quantify the linear relationship between the intrinsic (UMNSRS, MayoSRS, Bio-SimVerb and Bio-SimLex) and the extrinsic scores (CHEMD, BC2, AnatEM and PBA) [6].

Table 5.6 shows the correlation between the performances of representation models on various intrinsic evaluation datasets and the NER tasks. When compared to different benchmarks, the correlations between our datasets and downstream tasks are on par with or notably higher than the ones in UMNSRS and MayoSRS. The result suggests that our datasets can better predict the performance in NER, as compared with other intrinsic evaluation standards in biomedical NLP. Nevertheless, we find that there is no statistically significant correlation on any dataset (two-tailed t-test with $alpha = 0.05$). A possible reason is that the experiment involves only a limited number of data points, and only a large effect can be statistically significant.

[6] Spearman's ρ has been more natural for word similarity ranking comparisons given they are all cosine scores and we only want to compare the ranking. In contrast, here, using Pearson's r allows us to directly measure how well the performance metric in extrinsic tasks correlates with the intrinsic performance one.

Next, we compute the same performance-correlations using a set of SG models with different context window sizes (other hyper-parameters are kept default). The scores for individual tasks and their correlations are shown in Table 5.7 and Table 5.8 respectively.

With the same model architecture but different context window sizes, most extrinsic scores (right 4 columns of Table 5.7) have a performance peak with a narrow window (e.g. win= 1), followed by a gradual decrease when window size increases. The results in Table 5.8 show that our evaluation scores correlate better with downstream tasks than all other available intrinsic evaluation datasets. Although we only test on nine models, we observe two significant positive correlations in Bio-SimLex (CHEMD and AnatEM). Notably, UMNSRS and MayoSRS show a negative correlation with all NER tasks. Similar patterns are previously observed in Chapter 3 when comparing these scores using representation models trained with other corpora including PMC. They suggest that datasets such as MayoSRS emphasize modelling topical relatedness rather than similarity, which is learned better by a representation model with a larger context window. Nevertheless, tasks such as NER rely more on the modelling of similarity such as co-hyponymy, which is typically captured better with a narrow context window [Turney, 2012]. This disagreement in emphasis may lead to negative correlations between the intrinsic and extrinsic scores, as shown in Table 5.8. In contrast, we emphasized modelling relatedness and similarity separately during the annotation phase of Bio-SimLex and Bio-SimVerb. Annotators were instructed (with clear case examples) to give low scores to related but dissimilar word pairs, and this design leads to a higher correlation with extrinsic tasks in our experiments. Therefore, our datasets capture some properties of word similarity and relatedness that can predict performance at extrinsic tasks. Furthermore, Bio-SimLex shows a better correlation with extrinsic performance than Bio-SimVerb. One possible explanation is that the extrinsic tasks we considered in this experiment are NER, where performance is closely related to the quality of noun representations. More importantly, these results confirm our hypothesis that evaluating the qualities of the representation models separately for various word-types (e.g. verbs) provides insight into how they individually contribute to extrinsic performance.

Win Size	UMN-rel	UMN-sim	Mayo	Bio-SimVerb	Bio-SimLex	CHEMD	BC2	AnatEM	PBA
1	0.5317	0.5759	0.5551	0.4594	**0.7294**	**81.51**	70.06	82.16	65.34
2	0.563	0.6144	0.6238	**0.4696**	0.7207	81.44	70	**82.21**	65.51
4	0.5768	0.6247	0.581	0.464	0.7188	81.5	70.04	82	**65.75**
5	0.5767	0.6271	0.5744	0.4638	0.7151	81.37	**70.20**	81.32	65.16
8	0.582	0.6377	0.5975	0.4611	0.7086	81.24	69.56	80.99	65.53
16	0.5888	0.6431	0.6123	0.4667	0.7034	81.02	69.39	80.72	64.78
20	0.5896	0.6418	0.6319	0.4584	0.7031	81.12	69.62	80.49	65.19
25	**0.6018**	**0.6489**	0.6188	0.4519	0.7004	81.07	69.93	80.92	65.14
30	0.6007	0.6457	**0.6486**	0.4502	0.7043	80.71	69.2	81.03	64.79

Table 5.7 Intrinsic (left 5 columns, in ρ) and extrinsic scores (right 4 columns, in F-score) of the biomedical representation models trained using different window sizes.

	CHEMD	BC2	AnatEM	PBA
Bio-SimVerb (ours)	0.63	0.36	0.42	0.40
Bio-SimLex (ours)	**0.83***	**0.66**	**0.92***	**0.59**
Baseline				
UMN-rel	-0.78*	-0.56	-0.78*	-0.46
UMN-sim	-0.73	-0.57*	-0.81	-0.42*
MayoSRS	-0.78	-0.69	-0.54*	-0.47*

Table 5.8 Pearson's correlation between word-similarity/Bio-SimVerb and Bio-SimLex scores and the NER tasks evaluated on biomedical representation models trained with different window sizes (**Bold**: best scores, *: statistically significant)

5.4.4 Comparison with general-domain datasets

We have shown that our resources capture some properties (e.g. word semantics) that can predict performance in biomedical NER. These properties are expected to be domain-dependent, which suggests that it should be more effective to evaluate with in-domain datasets to predict performance for biomedical tasks. To study this, we use five representation models (detailed in Section 5.3), benchmarked on general-domain datasets (SimVerb and SimLex), and evaluate their performance-correlation on our datasets and biomedical tasks.

Table 5.9 shows the correlation between intrinsic and extrinsic scores for general-domain representation models. Most scores for general-domain datasets (SimLex and SimVerb) correlate negatively with biomedical NER tasks. Due to domain-specificity, the properties that SimVerb and SimLex measure generally do not reflect how well a representation model will perform in biomedical tasks, and may even give contradictory indications. Bio-SimLex achieves the best results also in this evaluation and shows a positive correlation with performance in BC2 and PBA despite measuring out-of-domain representation models. (In interpreting these results, it should be noted that none reaches statistical significance.)

To summarize, Bio-SimVerb and Bio-SimLex are better predictors of performance in biomedical NER than other in-domain datasets (UMNSRS, MayoSRS) and general-domain datasets (SimLex, SimVerb). We observed moderate to high positive correlations between performance on our datasets and in biomedical NER, which are consistent across corpora and different models as well as within the same model architecture with different windows sizes. Although it is possible to use our datasets to evaluate general-domain representation models, the results indicate that they are most effective in the evaluation of biomedical domain representation models.

	CHEMD	BC2	AnatEM	PBA
Bio-SimVerb (ours)	-0.38	-0.18	-0.47	-0.22
Bio-SimLex (ours)	**0.00**	**0.23**	**-0.09**	**0.18**
Baseline				
SimVerb	-0.31	-0.09	-0.41	-0.12
SimLex	-0.36	-0.20	-0.49	-0.19

Table 5.9 Pearson's correlation between general-domain datasets/Bio-SimVerb and Bio-SimLex scores and the NER tasks evaluated on general-domain representation models benchmarked in SimVerb and SimLex. None of the scores are statistically significant. (**Bold**: best scores)

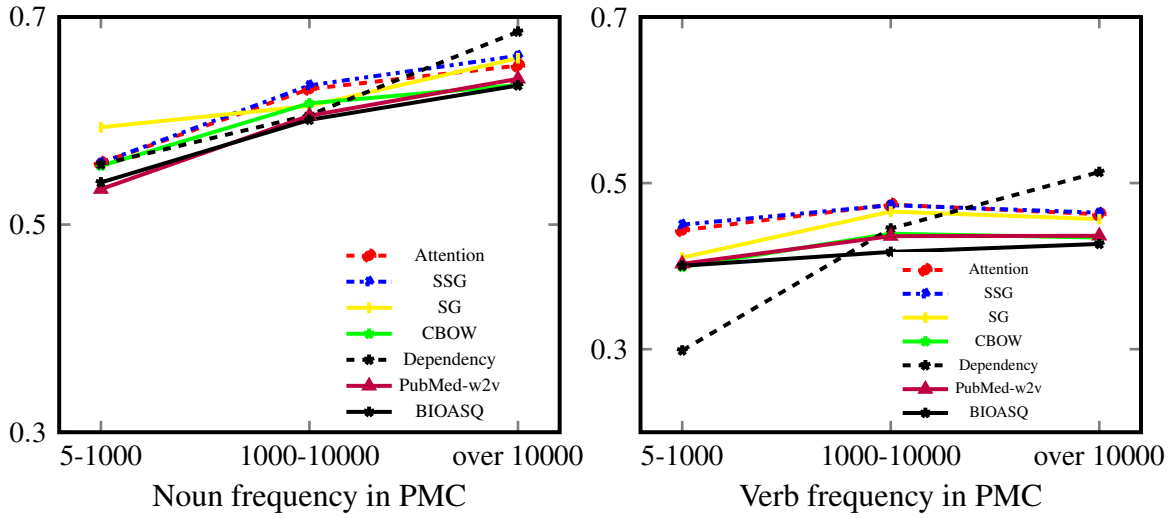


Fig. 5.1 Subset-based evaluation (y axis unit: ρ) for Bio-SimLex (left) and Bio-SimVerb (right), where subsets are created based on the word-frequency in PMC. To be included in each group it is required that both words in a pair are in the same frequency interval (x axis)

5.4.5 Subset evaluation

The extensive coverage and scale of Bio-SimVerb and Bio-SimLex enable model evaluation based on various criteria. In this section, we showcase two examples.

Frequency We first select word pairs based on their frequency of occurrence in PMC and form three groups, with 300-400 pairs in each group. Results for Bio-SimLex (left) and Bio-SimVerb (right) are shown in Figure 5.1. They suggest that the performance of all models improves as the frequency of the words in the pair increases. Since distributional models are data-driven, their qualities of capturing word-semantics are mainly governed by the word-frequency in the corpus.

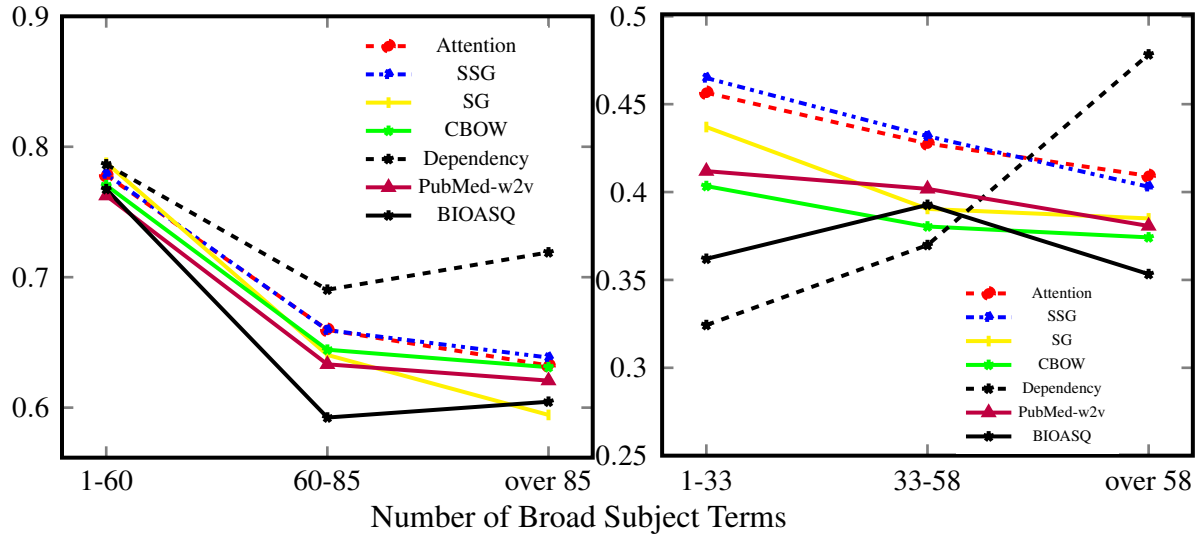


Fig. 5.2 Subset-based evaluation (y axis unit: ρ) for Bio-SimLex (left) and Bio-SimVerb (right). where subsets are created based on the word's number of unique Broad Subject Terms. A word can have multiple Broad Subject terms when it appears in journals of different areas in biomedicine. To be included in each group, it is required that both words in a pair are contained in the same Subject Term interval (x axis)

Broad Subject Terms In general, words with more diverse usage patterns (polysemy) are expected to be harder to learn with statistical models. To test this hypothesis, we divide the word pairs into three groups based on their numbers of Broad Subject Terms, which represent the sub-domains of text in which a word appears. Words that have more Broad Subject Terms appear in text across different areas of biomedicine and tend to have more diverse usage patterns compared to words used only in a single domain.

From Figures 5.2, we see a clear overall downward trend, suggesting that it is still a challenge for distributional models to capture the diverse usage patterns of words that appear across different domains. However, using additional information beyond corpus co-occurrence (e.g. dependency parsing) facilitates the learning of representation for such verbs, as reflected in the notable improvement for the dependency model seen in Figure 5.2 (right). Intuitively, dependency parses can provide discriminative context to facilitate representation learning: for example, two verbs are similar if they share similar nominal subjects (nsubj and nsubjpass). Nevertheless, our result shows that dependency parses do not contribute equally to the learning of noun and verb representations. Again, this supports our notion that representations of particular word types should be evaluated separately to better understand the type-specific properties learned by different models.

Frequency Subset	Bio-SimVerb	Bio-SimLex
low	0.9848	1.5621
medium	0.8059	0.6784
high	1.2352	1.0237
average	1.009	1.088

Subject Subset	Bio-SimVerb	Bio-SimLex
low	0.8941	1.2395
medium	0.9084	0.7585
high	1.25	1.1204
average	1.018	1.039

Table 5.10 Average standard deviation of ratings per subset (bold) by the word-frequency (left) and the number of Broad Subject Term (right). We use low, medium and high to label subsets for brevity. Range values of corresponding subsets can be found in Fig 5.1 and Fig 5.2.

Human Agreement Since distributional models are sensitive to word-frequency and the diversity of usage patterns, we also examine if these factors affect human perception of word similarity. In Table 5.10, we report the average standard deviation of ratings per subset by word frequency (left) and by Broad Subject Terms (right). That allows us to compare human agreement across subsets through the ratings of individual items in each subset. In general, the overall average standard deviations across all subsets are almost identical ($\approx$1.0). The subset where we find the highest deviation is the low-frequency subset of Bio-SimLex (left in Table 5.10). It is possible that annotators may not have been familiar with some rare words in Bio-SimLex, leading to higher variance in ratings.

5.5 Chapter summary

In this chapter, we have presented two new resources for the evaluation of word representation models: Bio-SimLex and Bio-SimVerb. These datasets allow researchers to investigate how humans and machines represent noun and verb semantics. Their sizes and coverage of concepts make it possible for the datasets to be used for comparing representation models in different areas of biomedicine. Furthermore, we have observed a positive correlation between the performance of biomedical representation models on Bio-SimLex and in biomedical NER. This indicates that our datasets can effectively measure properties that are relevant for performance in extrinsic tasks. We have also examined the impact of different representation learning approaches on nouns and verbs separately and observed that a single learning approach could not capture the semantics of all word types. To identify useful methods for learning type-specific representations, resources for the evaluation of individual word types, such as Bio-SimLex and Bio-SimVerb, are indispensable.

Chapter 6

A Neural Classification Method for Supporting the Creation of BioVerbNet

VerbNet, an extensive computational verb lexicon for English, has proved to be useful for supporting a wide range of Natural Language Processing (NLP) tasks requiring information about the behaviour and meaning of verbs, including word sense disambiguation [Brown et al., 2011], information extraction [Schmitz et al., 2012] and text mining applications [Lippincott et al., 2013; Rimell et al., 2013]. It is foreseeable that biomedical text processing and mining could benefit from a similar resource. However, while relatively well-developed resources are available for nouns in biomedicine (e.g. UMLS Metathesaurus, [Nelson et al., 2004]), verb-related resources are still lacking in both depth and coverage [Ananiadou and Mcnaught, 2006; Mondal et al., 2017; Tan, 2014; Venturi et al., 2009].

In this chapter, we will explore the application of neural representation for inducing verb lexicons in biomedicine. We will first describe how to fine-tune existing representation models to achieve this goal (Section 6.1). Then, will evaluate the utility of the induced lexicon in supporting biomedical NLP tasks (Section 6.2).

6.1 Creation of verb lexicon

Constructing verb lexicons manually is extremely labour-intensive. Instead, previous studies have shown that it is possible to automatically induce verb lexicons from both general [Barak et al., 2014; Joanis et al., 2008; Ó Séaghdha and Copestake, 2008; Vlachos et al., 2009] and biomedical texts [Korhonen et al., 2006, 2008; Sun, 2013]. Nonetheless, a vast majority of verb lexicon inductions rely heavily on feature engineering, which is time-consuming and requires added expert knowledge. Therefore, using hand-crafted features for verb lexicon

induction does not provide an optimal solution for lexicon acquisition in specific domains. However, in recent times, works which acquire general lexical resource on automatically-learned features (e.g. distributional word semantics) through neural networks are emerging [Vulić et al., 2017]. These sets of features are unsupervisedly induced from corpora, using neural representation learning models (a.k.a. neural embeddings) which have been fine-tuned to better capture verb semantics in the text. In the biomedical domain, there has been little work on the application and optimization of the neural representations for verb lexicon acquisition, partly attributed to the lack of an in-domain evaluation resource for verbs.

In the previous chapter, we have created a resource for the intrinsic evaluation of verb representations in the biomedical domain (Bio-SimVerb). It allows us to effectively measure properties that are relevant for learning verb-specific representations. In this chapter, we will investigate the potential of using verb-specific representations to develop a cost-effective, VerbNet-Style resource specifically aimed at describing verbs in the area of biomedicine. We propose an approach that can automatically identify contributive contexts for learning biomedical verb representations from large amounts of text without manual feature engineering. We will then apply this verb-optimized model on a small manual classification of biomedical verbs to expand it with new candidates using all the PubMed abstracts and the full articles in the PubMed Central Open Access subset as data.

6.1.1 Related work

Representation learning methods typically operate on two fundamental elements: *Words* and *contexts*. They represent a target word (*word*) using its neighbouring words (*contexts*). In literature, many models use the bag-of-words (BOW) contexts for learning the word representation, in which a fixed number of neighbours within the context window are used as the context of a target word. Nevertheless, using a fixed context window throughout the entire representation learning process implies that it is possible to miss important contextual features which appear outside the window. Conversely, expanding the window to capture all relevant words increases the level of noise from irrelevant words.

Hence, other types of contexts, such as dependency relations and symmetric patterns (e.g. *x* and *y*), have been proposed [Levy and Goldberg, 2014; Schwartz et al., 2015]. These studies show that representation learning requires different context configurations to produce improved results for each word-type (e.g. nouns or verbs). For example, Schwartz et al. [2015] reports that symmetric patterns (e.g. *x* or *y*) are essential to contexts for learning verb and adjective representations, whereas BOW is useful for learning noun representations.

Additionally, Vulić et al. [2017] proposes a framework for identifying the most useful (type-specific) contexts for learning representations for nouns, verbs and adjectives respectively.

These studies have only involved the general-domain text, and their results do not necessarily apply to the biomedical text. Additionally, in the biomedical domain, there has been little work on verb representation, partly attributed to the lack of an in-domain evaluation resource for verbs. With Bio-SimVerb we created in Chapter 5, we can now effectively measure properties that are relevant for learning verb-specific representations. Hence, in this chapter, we aim to identify the optimal dependency-based context configurations for learning representations of biomedical verbs, whose lexical characteristics can be different from the general-domain ones. We also extend the usefulness of neural representation by using our optimized models as features to induce a verb lexicon for supporting NLP and text mining in biomedicine. In the next section, we will describe its design and construction.

6.1.2 Dataset design

The design of our lexicon consists of two parts: First, we apply the recent method by Vulić et al. [2017] to identify best contexts for learning biomedical verb representations. The method, based on the Skip-gram model with negative sampling (**SGNS**), has produced successful results in the general domain but has not previously been applied to specialised domains such as biomedicine. It involves first creating a context configuration space based on dependency relations between words, followed by applying an adapted beam search algorithm [Pearl, 1984] to search this space for the verb-specific contexts, and finally using these contexts it creates verb representations.

Next, the optimized representation is used to provide word features for building a verb lexicon. This is obtained by expanding the small manually developed VerbNet-style classification of 192 biomedical verbs by Korhonen et al. [2006] (referred to as **Korhonen-VN** henceforth) with 957 new candidate verbs. The candidate verbs are chosen from the Bio-SimVerb we created in Chapter 5; verbs were selected based on their frequent occurrence in biomedical journals across 120 subdomains of biomedicine (as categorized by Broad Subject Terms [NLM, 2017]). It ensures the wide-coverage of verb classification ideal for the development of our lexicon. We then use the Nearest Centroid Classifier to connect the new candidates to an appropriate class in Korhonen-VN. The resulting classification provides 1,149 verbs assigned to the 50 classes in the original resource. It lists, for each verb, the most frequent dependency contexts that reflect their syntactic behaviour along with example sentences.

Context selection

Our aim is to fine-tune the learning of verb representation so that it can be used to build a verb lexicon. For this, we first identify contexts contributing to biomedical verb representation learning. We use the Stanford typed dependencies (DEPS, [De Marneffe and Manning, 2008]) as contexts for selection. It is because, first, DEPS can help representation models learn lexical information beyond the BOW context window and, second, they can provide a natural grouping of related words [Vulić et al., 2017]. For example: (*contain, glucose_dobj*) and (*generates, radiation_dobj*) which share the same dependency *dobj* can be grouped into *dobj* bag (referred to as context bag henceforth). In the next section, we describe how we construct these context bags.

Creation of context bags

We organized the dependency-parsed corpus for training representation in the form of (*word, context*) pairs, as in the work of Levy and Goldberg [2014]. *Word* is the target word for training the representation model whereas *context* stands for its corresponding context elements in text (e.g. dependency relations and the head of the dependent word). Consider the following as an example: the pair (*modulator, efficient_amod*) denotes a target word *modulator* with an adjectival modifier (*amod*) context: *efficient*. Given the dependency-parsed corpus, we break it down into individual context bags based on the dependency relation of each (*word, context*) pair. Hence, the context bag *dobj* consists of pairs such as *(regulate, cells_dobj)* or *(fire, neuron_dobj)*. We follow the same procedure as Vulić et al. to process the context bags. First, Prepositional and Conjunction relations are collapsed. Hence, all pairs with (*prep_x*) or (*conj_y*) such as (*prep_in*) and (*conj_or*) will be merged into the context bags (*prep*) and (*conj*) correspondingly. Secondly, similar dependencies (i.e. those at the bottom two levels of each dependency type in the Stanford dependency hierarchy) are merged. For example, direct (*nsubj*) and indirect subjects (*nsubjpass*) are merged into the context bag (*subj*). Thirdly, infrequent pairs and uninformative dependencies are removed (e.g. *punctuation*). A *context configuration* denotes a set of individual context bags used for training representation models. We call a configuration consisting of M individual context bags a *M-set configuration* [1]. Examining every possible context configuration is computationally expensive when there are many context bags. For example, assessing all contexts in a 10-set configuration (i.e. 10 context bags) would involve training $\sum_{k=1}^{10}(10C_k) = 1023$ different representation models [2]. In this case, we train and evaluate the representation

[1] For example, a *3-set configuration* will consist of 3 individual context bags such as *comp*, *conj* and *prep*

[2] A short formula for the summation of combination is: $2^{10} - 1 = 1023$

on 10 initial set of contexts and recursively consider smaller and smaller sets of them (from 10 to 1 set). We aim to improve the representation without exhaustively evaluating all possible combinations. To achieve this, we apply the context selection framework proposed by Vulić et al. [2017], which uses a beam search style selection to reduce the numbers of visited configurations. We will describe the details in the next section.

Configuration search

We implement the framework for context selection as proposed by Vulić et al. [2017]. First, we filter contexts that are uninformative for learning verb representation. For example, the *nn* bag denotes contexts linked from a noun to its noun pre-modifier. It is likely to be useful for learning noun representations, but not verb representations. Hence, when evaluating the quality of verb representation trained solely with the *nn* bag, we expect its score will be low. To filter uninformative contexts, we first train a set of representation models with every context bag we obtained from the dependency-parsed corpus, and evaluate them individually with Bio-SimVerb, the verb similarity gold standard we previously created in Chapter 5. A threshold score of $\rho \geq 0.2$ is used to filter uninformative contexts. Consequently, we use seven context bags as the initial configuration in our experiments. They are: *comp*, *conj*, *prep*, *pcomp*, *rel*, *subj* and *obj*. Vulić et al. [2017] suggest this step can effectively remove less relevant contexts at a minimal cost to accuracy.

After constructing the initial context configuration, the search algorithm starts from the full *M*-set configuration and tests *M(M-1)*-set configurations in which one individual bag is removed at a time to generate each such configuration. The algorithm narrows down the search by keeping only those sets of configurations which outperform the original M-set configuration. It continues searching over lower-level *(M-1)*-set configurations until it reaches the lowest level or when no further improvements over its original configurations are found.

Using this context selection framework, the search for the optimal configuration for verbs is reduced to only 27 context configurations out of 127 possible configurations ($2^7 - 1 = 127$). It includes seven 1-set configurations (i.e. individual context bag) plus twenty other configurations. After we identify the optimal context configuration for verbs, we train a representation model with this configuration. This model will be used for constructing an initial candidate grouping for our verb lexicon. We describe our construction in the next section.

Verb classification

As described earlier, we expand Korhonen-VN with a list of new candidate verbs selected from Bio-SimVerb. We use Bio-SimVerb as a source for candidate verbs for multiple reasons: first, it contains verbs that have been manually validated by domain experts, chosen based on their common usages in biomedical text. It avoids the problem of including overly general verbs such as 'have' and 'be' or too specific verbs such as 'x-ray'. Second, these verbs have been sourced from journals across 120 sub-domains of biomedicine (as categorized by Broad Subject Terms [NLM, 2017]), ensuring extensive coverage over different areas, which is essential as our methodology is ultimately aimed at supporting the creation of a large-scale VerbNet-style resource. Furthermore, since we evaluate our models against Bio-SimVerb, we expect that our optimized model can best capture the syntactic and semantic properties of verbs in Bio-SimVerb. Finally, to connect new candidates to a class in the existing verb resource, we use the Nearest Centroid Classifier. It represents each class by the centroid of its member verbs in vector space (from our optimized representation) and connects the new candidates to their nearest class centroids (in terms of Euclidean distance).

To investigate the suitability of our classification methodology for facilitating the cost-effective creation of a large-scale verb resources, we considered human evaluation. We will now describe it in detail.

Human evaluation of verb classes

Since our aim is to investigate the suitability of our classification methodology for facilitating the creation of a verb lexicon, we use human experts (two linguists and two biologists) to evaluate the new member verbs and possible novel classes in the sample of a classifier output. Following well-established practices in related works [Majewska et al., 2018; Sun, 2013], the experts' task is to determine whether the new member verbs within each verb class are similar enough in terms of their meaning and syntactic patterns to the existing verbs in the original classification. This is done to determine if each verb is a legitimate member of the verb class. Whenever this is the case, the method has accurately learned correct classification. When this is not the case, the verbs are examined further for potential discovery of new subclasses to be included in the original classification. When verbs are clearly misclassified, they are excluded or re-assigned to other classes as agreed by the experts.

For this evaluation, the experts followed guidelines specifically developed for the purpose (can be found in Appendix A). The data provided for the experts include the original class names and member verbs from Korhonen-VN, the new member verbs from the classifier output and the set of 10 most frequent dependency contexts for each verb. For example, all

verbs from the Class 1.2 are labelled as 'Verb of affection', the class consists of member verbs such as *modulate* and *regulate* which are used to describe events that have an effect on entities. The dependency context of *regulate* as in the sentence *Dox could effectively regulate bFGF expression* is denoted as (*subj#obj*). Thirty sentences, three per the ten most frequent dependencies of each verb, are also provided along with the dependency information to demonstrate how each verb is used in context.

6.1.3 Dataset construction

Word representation

Data The dependency-parsed corpus is compiled from the pre-processed PubMed Central Open Access subset (PMC) and PubMed abstracts which are distributed by Hakala et al. [2016]. POS tags and tokens in this resource are generated using the BLLIP constituency parser [Charniak and Johnson, 2005] trained on a biomedical corpus [Mcclosky, 2010] [3]. The resource covers over 26M abstracts and 1.4M full articles with more than 388M parsed sentences. We filter out words that appear fewer than 100 times in the text, as suggested in the work of Levy and Goldberg [2014]. Consequently, the corpus consists of approximately 27 million word types.

Model In this experiment, we use the popular Skip-gram model with negative sampling architecture (**SGNS**) to train the word representations. Levy and Goldberg [2014] have developed a tool which allows SGNS to learn representations from dependency-parsed contexts formatted as (*word, context*) pairs. All representation models used in this experiment are trained with vector dimension (*d=300*). Similar settings can be found in other studies [Schwartz et al., 2015; Vulić and Korhonen, 2016]. The baseline model we used is a SGNS trained with all dependency contexts in the corpus (**DEP-ALL**), a SGNS model trained only with the seven verb-related contexts (**POOL-ALL**) we identified in section 6.1.2 (i.e. contexts with evaluation scores $\rho \geq 0.2$ on Bio-SimVerb) and a standard SGNS trained with bag-of-words contexts (**BOW**) using the **word2vec** tool [Mikolov et al., 2013a]. These models are used to compare against other representation models with different context configurations. The best-performing model (as evaluated with Bio-SimVerb) is then used to build the prototype of our verb lexicon that is validated and corrected manually.

Evaluation The Bio-SimVerb (a word similarity evaluation dataset we created in Chapter 5) is used as the gold standard to measure the quality of our verb representation models. It consists of 1,000 verb pairs whose degree of similarity are ranked by human judges. The similarity ranking of a given representation model is computed as the cosine similarity of the vectors of these verb pairs. Following the standard evaluation protocol, we compare the similarity rankings produced by humans and by individual models on those 1,000 verb pairs using the Spearman's ρ correlation. A higher correlation value implies a better model in capturing verb semantics in the text.

[3] The description of corpora is provided in Section 5.2.1

Index	Class name	Subclass name	Example members
2.2.1	Biochemical events	Biochemical modification	dephosphorylate, phosphorylate
4.1.3	Experimental procedure	Label	stain, label, immunoblot, probe fix
4.2.0		Precipitate	coprecipitate, coimmunoprecipitate, precipitate
9.1.1	Report	Examine	assess, evaluate, estimate, examine, explore analyze
9.1.2		Establish	establish, test, investigate
9.2.1		Presentational	argue, hypothesize, conclude, reason, note, speculate, assume
10.1.1	Perform	Quantitate	quantify, quantitate, measure, monitor
11.0.0	Release	Release	release, detach, excise, dissociate
12.0.0	Use	Use	utilize, employ, exploit
14.0.0	Call	Call	name, designate
16.0.0	Appear	Appear	become, occur, seem

Table 6.1 Example gold standard classes and class members from Korhonen et al. (2006)

Verb resource (Korhonen-VN)

Korhonen et al. [2006] manually developed a VerbNet-style gold standard for verb classification in biomedicine containing 192 verbs organised into a class taxonomy of 50 fine-grained classes for biomedical verbs (Examples are shown in Table 6.1). To the best of our knowledge, it is the only biomedical resource of this type. We use this resource as a starting point for the creation of a supervised approach intended to facilitate the development of a verb lexicon. Essentially our goal is to expand Korhonen-VN by automatically connecting new candidate verbs to existing verbs based on their Euclidean distance found in the vector space of an optimized representation model.

Model	Spearman's ρ
Baseline	
BOW (win=5)	0.4664
DEP-ALL	0.4323
Configurations: Verb	
POOL-ALL	0.4724
conj+obj+pcomp+prep+rel+subj	0.475
conj+obj+prep+rel+subj**(Best)**	**0.4889**
conj+obj+pcomp+prep+subj	0.4578
conj+obj+pcomp+rel+subj	0.4478
conj+obj+pcomp+prep+rel	0.4406
conj+obj+prep+subj	0.4611
conj+obj+rel+subj	0.4572
conj+obj+prep+rel	0.442
comp+obj+pcomp+prep+rel+subj	0.4376
comp+conj+obj+prep+rel+subj	0.4762
comp+conj+obj+pcomp+prep+subj	0.4655
comp+conj+obj+pcomp+rel+subj	0.4583
comp+conj+obj+pcomp+prep+rel	0.4413
comp+conj+obj+prep+subj	0.4635
comp+conj+obj+rel+subj	0.4592
comp+conj+obj+prep+rel	0.442
obj+pcomp+prep+rel+subj	0.4446
obj+prep+rel+subj	0.441

Table 6.2 Performance on Bio-SimVerb (in ρ) using representations learned with different context configurations. BOW denotes a basic SGNS learned with bag-of-words context with the context window size 5. DEP-ALL denotes a configuration where no filtering of contexts are used. POOL-ALL denotes a configuration where all individual context bags from the verb-related pools are used. "Best" identifies the best-performing configuration found.

6.1.4 Results

Representation learning

We examine whether different context configurations can improve the quality of verb representation when evaluated against human judgments on a verb similarity task (Bio-SimVerb, as measured on ρ points). Results are shown in Table 6.2. In general, selecting an optimal context configuration for verbs gives better performance. From Table 6.2, there is an apparent difference (5 ρ points) between models trained with and without context selection: While an

evident improvement (4 ρ points) can already be found when we pool only contexts that are useful for verbs (POOL-ALL, detail in Section 6.1.2) from the generic corpus (DEP-ALL). A further selection among these verb-related contexts yields additional improvements (1 ρ point). Overall, the model trained with the best context configuration is approximately 2 ρ points over the best baseline. The results provide us some linguistic insights on which contexts are contributive to the learning of biomedical verb representations. For example, two verbs are similar if they are used with similar subjects *subj*, objects *obj* as well as pronoun phrases in a relative clause *rel*. Also, semantically similar verbs are commonly connected by the conjunction likes *and* (e.g. walk *and run*). In this study, we observe that identifying verb-specific contexts is valuable for learning verb representations.

Automatic verb classification

To classify verbs into semantic groups, we run a Nearest Centroid Classifier on top of the verb-specific representations, using vector dimensions as features for learning verb classes. The classifier is first trained using verbs in Korhonen-VN. It then connects new verbs to classes based on their Euclidean distance. Consequently, 957 verbs are classified into 50 classes.

Human validation of verb classes

In order to evaluate the output of the classifier, we employed four experts, two linguists and two biologists with at least a postgraduate level of training in their subject areas. The experts first performed the validation of selected classes individually according to the guidelines (included as Appendix A), and then consulted and discussed their validations in each domain-specific pair and in linguist-biologist pairs. The 14 classes selected for validation were chosen at random from the classifier output so as to ensure that both the biomedical and the general scientific domains were represented, with 7 classes chosen per domain, each class consisting of 4-28 member verbs.

The experts were presented with written guidelines and the following materials: (1) a file including the verb classes, their original members from Korhonen-VN, and the new candidates to be reviewed (Table 6.3); (2) an Excel spreadsheet for recording the updated index of the class for each verb based on the manual revision of the class candidates, (3) 30 example sentences drawn from the corpus used in the experiment representing the 10 most frequent dependency contexts for each verb.

The guidelines instructed the experts to verify whether the new candidate verbs were similar in terms of their meaning as well as syntactic patterns to the existing member verbs

Index	Subclass name	Example members (from Korhonen-VN)	New candidates
1.1.2	Suppress	suppress, repress	downregulate, transactivate
7.1.0	Collect	harvest, select, collect	decide, pick, cultivate, procure, gather, choose, transfuse, prioritize, obtain
13.1.0	Encompass	encompass, possess, comprise, bear, span, harbor	overlie, display, hold, exhibit, cover, infest, belong, range
14.0.0	Call	call, name, designate	qualify, regard, rename, mention, request
1.4.0	Modify	modify, catalyze	hydroxylate, hydrolyze, methylate, deaminate, esterify, oxidize, detoxify, metabolize
4.1.3	Label	stain, label, immunoblot, probe, fix	supershift, assay, immunostain, tag, immunolabel, clone, postfix, digest, clamp, counterstain, buffer, electroblot, fluoresce, radiolabel, blot
11.0.0	Release	release, detach, excise, dissociate	reinsert, retract, disassemble, deacylate, extrude, remove, depolymerize, mobilize, lose, resect, separate
10.1.3	Conduct	perform, conduct	execute, undertake

Table 6.3 Example classes validated by experts

in the original classification. The 30 example sentences provided were meant to facilitate the review process by illustrating how a given verb is used in biomedical texts (keeping in mind that this may differ from its typical usage in the general language domain), i.e. the most common syntactic structures in which it appears. Based on the analysis of the semantic and syntactic behaviour of the new candidates with respect to the existing class members, the experts were asked to decide if each new candidate has been correctly assigned to a given class, or if not, whether it should be (a) reassigned to another class in the classification, (b) form a subclass within a broader existing class, or (c) should be moved to a new class altogether (along with some other misclassified verbs); or otherwise, if no appropriate class could be thought of, (d) whether it should be discarded as noise (i.e. a mistake by the classifier). Polysemous verbs were controlled, we chose verbs and classes that have less diverse usage patterns (i.e. verbs that have less than twenty Broad Subject Terms [4]. We further consult experts to assure that most of the chosen verbs have a dominant sense. Consequently, a given verb could only be assigned to a single class or subclass (i.e. soft clustering was not permitted).

[4]The description of Broad Subject Terms is provided in Section 5.2.1

	#New candidates	#Correct candidates	%Correct candidates	#Incorrect candidates	%Incorrect candidates
7.1.0 COLLECT	9	6	66.7	3	33.3
9.1.1 EXAMINE	21	19	90.5	2	9.5
9.3.0 INDICATE	11	10	90.9	1	9.1
10.1.3 CONDUCT	2	2	100	0	0.0
13.1.0 ENCOMPASS	8	6	75.0	2	25.0
14.0.0 CALL	5	4	80.0	1	20.0
16.0.0 APPEAR	19	16	84.2	3	15.8
General total	75	63	**83.9**	12	**16.1**
1.1.2 SUPPRESS	2	2	100	0	0.0
1.1.4 INACTIVATE	15	11	73.3	4	26.7
1.4.0 MODIFY	8	6	75	2	25
2.3.0 INTERACT	21	19	90.5	2	9.5
4.1.3 LABEL	15	11	73.3	4	26.7
8.3.1 TRANSPORT	19	17	89.5	2	10.5
11.0.0 RELEASE	11	10	90.9	1	9.1
Biomedical total	91	76	**84.6**	15	**15.4**
Total	166	139	**84.3**	27	**15.7**

Table 6.4 Results of class validation by experts, for seven general scientific (General) and seven biomedical classes (Biomedical), and across the two domains (Total)

Qualitative analysis

After having completed the validation task, the experts compared and discussed their analyses to come up with the final classification that they agreed on. The results of the validation are presented in Table 6.4.

The evaluation shows that over 83% of the new candidates generated across the two domains are valid class members, and in each of the 14 individual classes the majority of novel classifications are correct. From the total number of 166 novel candidates, 139 were judged as correct, which demonstrates that our automatic method can be used as a highly accurate starting point for the creation of a verb lexicon.

In two of the evaluated classes, 'Conduct' in the general domain and 'Suppress' in the biomedical domain, all of the novel classifications were marked as valid member verbs, while in four other classes - 'Examine' and 'Indicate' in the general domain and 'Interact'and 'Release' in the biomedical domain - over 90% of new candidates were judged as correct. The 'Conduct' class provides a good example of how our system accurately selects candidates that are semantically similar to the existing class members based on similarity of their syntactic behaviour: the original member verbs, *perform* and *conduct*, are provided with

new synonymous counterparts, *execute* and *undertake*. Analogous cases are found in the biomedical domain, e.g., in the 'Interact' class, a new candidate *collaborate* is a close synonym of one of the original class members, *cooperate*. What is more, our classifier picks up not only synonymous but also antonymous verbs as candidates for a given class, as seen in the biomedical domain (e.g. *downregulate - transactivate*). It is consistent with what has been observed in previous work on the manual semantic classification of verbs [Majewska et al., 2018], where human annotators were found to consistently group antonyms together as semantically similar. Despite representing the opposites of a meaning continuum, antonyms have almost identical distributions, and that paradigmatic similarity is what makes annotators judge them as semantically closely related.

An in-depth analysis of the candidate verbs by the experts sheds light on the strengths of the presented approach, as well as the error patterns and areas for future improvement. Overall, only 15.7% of new candidates were judged as incorrect across all 14 classes, with slightly more noise found in the general scientific classes (16.1%) than in the biomedical classes (15.4%). In the general language domain, the linguists identified between 0 to 3 incorrect candidates per class, whereas in the biomedical domain, the experts marked between 0 to 4 candidates per class as incorrect for the class in question, either judged as mistakes or as candidates for reassignment to another class.

Several recurrent reasons behind the erroneously classified verbs can be identified:

a. **Verbs share syntactic but not semantic similarity:** Examples of candidates which ended up in a given class purely through accidental syntactic similarity to the existing members are found, for instance, in the biomedical class 'Transport'. The two incorrect candidates identified, *tailor* and *generalize*, share the syntactic contexts of subj#obj (*The methods generalize earlier approaches...*), subj#prep (*This advantage did not generalize to the visual domain*), and subj#obj#prep (*We also generalize some known results from the real-valued case to the complex-valued one*) with the original class members (e.g. *Highly resistive wires* transmit *intracardiac electrograms*, *Occasionally these viruses* transmit *to other mammals*, *GPCRs* transmit *signals through heterotrimeric G proteins*). In the general scientific domain, examples of coincidentally parallel syntactic behaviour between new and original class members were noted, for instance, in the 'Collect' class: *decide* and *prioritize*, marked as noisy, share the syntactic contexts of subj#obj (*Future research should prioritize addressing symptoms...*) and obj#prep (*Should the surgeon decide on relaparoscopy...*) with *harvest, select* and *collect*.

b. **Parsing errors:** In some cases the syntactic contexts themselves were mistakenly identified as identical due to a parser error, which produced noisy candidates. For example, the verb *lie* got classified with the 'Appear' class members based, among others, on the shared subj#obj#prep context, exemplified by the phrase: *Thermal imaging as a lie detection tool at airports*, where 'lie' is a noun modifier of 'detection', both of which form a compound modifying the noun 'tool', rather than being a verb taking a noun object and a preposition. Or similarly, in the context subj#obj#prep *We review the technical challenges that lie ahead*, 'ahead' is mistakenly analyzed as the object rather than a preposition. Another type of error had to do with analyzing the particle 'to' as a preposition rather than an infinitive marker, as in the few cases of misidentified syntactic contexts such as *HIV and HCV seem to co-opt DDX3* as identical to subj#prep: *many interventions may vary between population groups*, or *(...) await for clinical applications*, which contributed to clustering dissimilar verbs such as *vary, await, pave* together in the 'Appear' class with *appear* and *seem*.

c. **Clustering loosely related verbs (rather than strictly semantically similar):** Another type of misclassification involves candidate verbs which are related to the existing class members but are dissimilar to them with respect to some meaning components or semantic properties identified as characteristic of the class in question. In the biomedical domain, examples of this kind of error are found in the 'Modify' class, where 8 new candidates are added: *hydroxylate, hydrolyze, methylate, deaminate, esterify, oxidize, detoxify, metabolize*. Out of these, the last two (*detoxify, metabolize*) were flagged as standing out from the rest, based on the fact that they describe processes on the cellular level, in contrast to the rest of member verbs referring to a specific chemical changing (i.e. terms pertaining to the chemical level). In the general scientific domain, examples of related but not strictly similar verbs added through looser association with the existing members include *optimize* and *understand* yielded for the 'Examine' class, or *cultivate* in the 'Collect' class.

In some cases, the new verbs judged as not valid were marked as candidates for reassignment to another existing class, or as members of a subclass or a new class altogether. An incidence matrix showing the class reassignments is presented in Appendix B. For instance, *exacerbate, aggravate* and *magnify*, found in the 'Inactivate' class, were highlighted as forming a separate cluster of similar verbs, while the verb *deacylate* found in the 'Release' class was reassigned to the 'Modify' class. In the general scientific domain, an example of reassignment involved verbs *display* and *exhibit*, considered better suited for the 'Indicate' class, within which four other candidates, *underline, underscore, highlight, emphasize*, were

marked as forming a subclass of 'underline'-type verbs. Such cases demonstrate the potential of the classification method for also discovering valid novel classes not in the original classification.

6.1.5 Discussion

The in-depth scrutiny of the new candidates shows that our automatic classification approach is highly accurate and thus likely to be very useful for extending the manual classification of biomedical verbs to a large-scale lexical resource. Although some human validation and filtering of the noise is necessary for the development of a fully accurate resource, the time and cost required for this are likely to be small in comparison with a fully manual development of such a large resource from scratch. The manual development of the original Levin classification [Levin, 1993] and VerbNet [Kipper-Schuler, 2005] required years of research efforts, although semi-automatic methods were used to facilitate their extensions too [Kipper et al., 2008]. Our qualitative analysis shows that despite being based purely on syntactic behaviour and combinatorial properties of verbs, the method also associates verbs in terms of their shared semantics, yielding classes of semantically similar and closely related members.

The error analysis reveals some areas of potential improvement. While the accidental syntactic parallels are a difficult problem to deal with (and have, in fact, been reported to challenge verb classification regardless of the clustering approach adopted [Sun, 2013]), errors from parsing could be addressed in the future via use of tools capable of dealing with the problem cases highlighted in the previous section. Misclassifications involving candidate verbs which are related to the existing class members but dissimilar to them with respect to some semantic properties are not necessarily an issue that needs to be addressed. Rather, such cases may actually demonstrate the potential of the method to hypothesize new classes and classifications for human validation and offer the means for subsequent refinement of the original classification. It is important because the original classification is, by no means, comprehensive and is likely to require further development as we scale it up to cover language in biomedicine.

6.2 Task-based evaluation

So far, we have described the creation of our lexicon from a verb-optimized neural representation and its evaluation based on human judgments. In this section, we will evaluate the utility of the verb lexicon in regards to improving downstream NLP tasks when used as features. For this, we first apply the *retrofitting* approach as proposed by Faruqui et al. [2015] to incorporate the verb class information (as obtained from our automatically-created lexicon) into the vector-space representation. The retrofitted-representation will then be used to provide features for text classification and relation classification tasks.

6.2.1 Related work

Lexical resources can be used to enrich representation models by providing them other sources of linguistic 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 type of 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 type of method incorporates lexical information into the vector representations as a post-processing procedure. The method *fine-tunes* the pre-trained word vectors to satisfy linguistic constraints from the external resources. The method can be applied to any off-the-shelf model without requiring large corpora for (re-) training as in the joint-learning models do. Among these methods *retrofitting* [Faruqui et al., 2015] is widely used; given any (pre-trained) vector-space representations, the goal of retrofitting is to bring closer words which are connected via a relation (e.g. synonym) in a given semantic network or lexical resource (i.e. linguistic constraints). For example, Yu et al. [2016] retrofit word vector spaces of MeSH terms by using additional linkage information from UMNSRS to improve the representations of biomedical concepts. Additionally, building upon retrofitting, Lengerich et al. [2018] generalize retrofitting methods by explicitly modelling individual linguistic constraints that are commonly found in health/clinical-related lexicons (e.g. causal-relations between diseases and drugs).

In theory, the joint-learning models could be as effective (or better) than the ones produced by fine-tuning distributional vectors. However, the performance of joint-learning models has

not surpassed that of fine-tuning methods. [5] Furthermore, the joint-learning objectives are usually model-specific and are tailored to a particular model making them difficult to be used in other methods. In this work, we will use retrofitting to incorporate our lexical features into the word representations.

6.2.2 Methodology

We base our retrofitting method on the one proposed by Faruqui et al. [2015]. Given any pre-trained vector-space representation, the main idea of retrofitting is to pull words which are connected in relation to the provided semantic lexicon closer to the vector space. The main objective function to minimize in the retrofitting model is expressed as:

$$\sum_{i=1}^{|V|} \left(\alpha_i \left\| \vec{v}_i - \vec{\hat{v}}_i \right\| + \sum_{(i,j)\in S} \beta_{ij} \left\| \vec{v}_i - \vec{v}_j \right\| \right) \tag{6.1}$$

where |V| represents the size of the vocabulary; $\vec{v}_i$ and $\vec{v}_j$ corresponds to word vectors in a pre-trained representation, and $\vec{\hat{v}}_i$ represents the output word vector, which is fine-tuned with the lexical constraints. S is the input lexicon represented as a set of *linguistic constraints*. In our case, they are pairs of word indices, denoting the pair-wise relations between member verbs in each 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 in the lexicon. α_i and β_{ij} are pre-defined values that control the relative strength of associations between members. The ranges of α and β is 0-1. A large α would constrain the retrofitted vectors to be as similar as the initial ones. It aims to preserve the high-quality semantic content as presented in the initial vector space, as long as this information does not contradict the linguistic constraints to be retrofitted. On the other hand, the retrofitting process will be unstable if β is too large and will be slow if it is too small. Here, we follow the default settings as stated in the authors' work and use $\alpha = 1$ 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 [Faruqui, 2015]. We used their implementation in the current work.

Baseline word representation

Based on the parameter selection experiments covering corpora, model architectures and hyper-parameters (Chapter 3), we selected the best-performing options for learning biomedi-

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

cal representations. Here, we include the model trained with those options as the baseline for our comparative evaluation against the retrofitted models. Table 6.5 shows the options' training settings: [6]

Parameters	Values
Corpus	PubMed
Architecture	Skip-gram
neg	10
dim	200
alpha	0.05
samp	1e-4
win	2
min-count	5

Table 6.5 Settings selected for comparative evaluation

Semantic lexicons

Our automatically-created lexicon is generated by extending **Korhonen-VN**. Hence, we include both lexicons and compare their utilities in improving the representation models. Both lexicons are organized in a hierarchical form, which consists of three levels with 16, 34, 50 verb classes on each level correspondingly. Table 6.6 shows the linguistic constraint counts under each class as derived from the two lexicons. When retrofitted against the three top levels, those member verbs at each sub-class are merged with its upper-class, as in the work of Faruqui et al. [2015].

	#Verb pairs	
	Korhonen-VN	Our lexicon
16-classes	1,774	96,998
34-classes	638	54,063
50-classes	376	50,104

Table 6.6 Linguistic constraint counts under each class as obtained from the Korhonen's resource and our automatically-created lexicon. Total number of verbs (Korhonen-VN: 192, our lexicon: 1,149).

[6] One can refer to Chapter 3 for the description of individual parameters (Section 3.3.3)

6.2.3 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. We will now describe them in further detail:

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. It was introduced in the work by Weinberg and Hanahan [2000] and has been widely-used in biomedical NLP for many systems and works, such as 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 from PubMed. The data consists of multi-labelled documents and sentences using a taxonomy of 37 classes.

The Exposure taxonomy is an annotated dataset for the classification of text (documents or sentences) about chemical risk assessments, as introduced by Larsson et al. [2017]. It is related to the assessment of exposure routes (such as ingestion, inhalation, or dermal absorption) and human bio-monitoring (the measurement of Exposure Bio-markers). The taxonomy of 32 classes is divided into two branches: Bio-monitoring and Exposure routes. Table 6.7 shows the basic statistics for each dataset. [7]

	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 6.7 Summary statistics of the Hallmarks of Cancer (HOC) and the Exposure Taxonomy (EXP)

[7] We refer the interested readers to Baker [2018] (Table 3.12 and 3.13) for the detailed description and distribution of the classes.

Model The model follows the convolutional neural network model (CNN) proposed by Kim [2014]. An implementation of this algorithm on HOC and EXP has been published by Baker and Korhonen [2017]. We use their implementation in our experiment. The model input 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 for predicting labels for text. Model hyper-parameters and the training set-up are summarized in Table 6.8:

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 6.8 Hyper-parameters used in Baker and Korhonen [2017]

Performance is evaluated using the standard precision, recall, and F-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. Besides, to address over-fitting in the CNN, we follow the authors' early stopping approach: testing only the model that achieved the highest results on the development set.

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 is installed with the Turku Extraction System [Björne, 2014]. It is parsed with 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 6.9 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 6.9 Summary statistics of the Chemical-Protein interaction dataset (CHEMPROT)

Model 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-score metrics of the labels in the model. Classification is performed as multilabel classification where each example may have 0–n positive labels. Model hyper-parameters and the training set-up are summarized in Table 6.10.

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 6.10 Hyper-parameters used in Björne and Salakoski [2018]

To account for variance in neural networks due to their random initialization, we follow the ensembles settings as used in Björne and Salakoski [2018]'s work. We train 20 models and take the n-best ones (n=5), ranked with their F-score on the development set, and use their averaged predictions. The ensemble predictions are calculated for each label as the average of all the models' predicted confidence scores. Furthermore, 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. Conversely, to focus on the effect of our lexicon on biomedical representations, we experiment with word representations induced on biomedical

texts, diverging from the authors' work which use the embeddings from Pyysalo et al. [2013a] that is induced on a combination of biomedical and general-domain data (PubMed, PMC and Wikipedia texts).

6.2.4 Results

We evaluate the utility of our automatically-created lexicon by improving the word representations (through retrofitting). We compare the performance of the baseline with the retrofitted representation models by measuring their precision (P), recall (R), and F-scores in text classification and relation classification when used as features.

For text classification, Table 6.11 and Table 6.12 show the micro-averaged scores for HOC and the EXP respectively. Each table shows performance on document and sentence-level classification (as columns) with different semantic lexicons (as rows). For relation classification, Table 6.13 shows the micro-averaged scores for CHEMPROT. Best results are shown in bold and statistically significant scores are shown with an asterisk. All statistical tests are done using a two-tailed t-test with $\alpha = 0.05$. We first describe experiments measuring improvements from the retrofitting method, followed by comparisons to using different sets of lexicons during retrofitting.

Retrofitting

We use Eq. 6.1 to retrofit word representations using linguistic constraints derived from verb lexicons. Overall, the retrofitted models show improvements in most tasks.

For text classification, better scores are found in three out of the four cases. From the results of HOC in Table 6.11, all retrofitted models outperform the baseline in F-score, which is contributed by a substantial improvement in recall particularly at the document level where there is a 15 point increase over the baseline. In total, five out of twelve improvements reported are statistically significant. The results for EXP in Table 6.12 are more mixed. At the document level, all retrofitted models achieve a slight F-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 amounts 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 6.13, there is about 1.5 point (F-score) increase over the baseline, and half of the improvements reported are significant.

The results from both tasks suggest that the semantic classes provided by verb lexicons improve performance over the raw verb features.

Semantic lexicons

When compared among the tested-lexicons, we find that our automatically-created lexicon clearly gives a better improvement to the baseline representations over the Korhonen-VN in all evaluated tasks. One possible reason is that our lexicon has a notably larger dataset size in comparison to the Korhonen-VN (see Table 6.6), 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, yet, the degrees of generalization needed may vary from task to task. When experimenting retrofitting with different levels of verb classes, we observe a notable difference (1-2 points in F-score) between models retrofitted with the coarse-grained level of 16-classes and the fine-grained level of 50-classes. For text classification (Table 6.11 and Table 6.12), on the document-level classifications (in both datasets), models appear to benefit by a finer-grained classification of 50-classes, whereas on the sentence-level classifications, a medium-level of generalization (34-classes) seems optimal. In relation classification (Table 6.13), the best result is also obtained with a medium-level of generalization (34-classes).

	Document classification			Sentence classification		
Lexicon	P	R	F_1	P	R	F_1
Baseline (no lexicon)	77.8	51.7	62.1	56.8	30.7	39.9
Korhonen-VN						
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
50-classes	74.9	59.2	66.2	48.4	35.2	40.7*
Our lexicon						
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 6.11 Performance results for the Hallmarks of Cancer (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. Its scores are adopted from Baker and Korhonen [2017]. All figures are micro-averages expressed as percentages (Bold: the best score, *: statistically significant)

	Document classification			Sentence classification		
Lexicon	P	R	F_1	P	R	F_1
Baseline (no lexicon)	89.5	87.1	88.3	66.2	62.8	**64.5**
Korhonen-VN						
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
Our lexicon						
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 6.12 Performance results for the Exposure Taxonomy (EXP) when different sets of lexicons are used for retrofitting the baseline model. Baseline denotes a Skip-gram model generated with our optimized training settings. Its scores are adopted from Baker and Korhonen [2017]. All figures are micro-averages expressed as percentages. (Bold: the best score for a task, *: statistically significant)

Lexicon	P	R	F_1
Baseline (no lexicon)	76.9	63.5	69.5*
SOTA (no lexicon)	75.1	65.1	69.7
Korhonen-VN			
16-classes	76.5	64.6	70.1
34-classes	78.2	63.8	70.3*
50-classes	76.5	65.0	70.3*
Our lexicon			
16-classes	76.3	65.2	70.3
34-classes	77.5	65.6	**71**
50-classes	76.2	65.9	70.7*

Table 6.13 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 Pyysalo et al. [2013a]s' embeddings. All figures are micro-averages expressed as percentages. (Bold: the best score for a task, *: statistically significant)

6.2.5 Discussion

The task-based evaluations suggest that our lexicon, as induced from a verb-optimized representation, can be a useful resource in supporting 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, our lexicon appears to have captured some of these 'topic-related' properties. In HOC, we notice that some high frequent verbs appeared in documents relating to the topic: *Sustaining proliferative signaling*, e.g. *proliferate* and *grow* also share the same classes in our automatically-created lexicon. Similarly, for exposure assessments documents describing air monitoring data in EXP, we can frequently see member verbs in the 'Proceed' class such as *inhale* and *breathe*.

Entities-relations described in the biomedical literature are often expressed in a predicative form where a trigger word (mostly a verb) connects two or more entities, and a range of 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 e.g. the 'Suppress' class in our lexicon. 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. Additionally, our lexicon covers more verbs than Korhonen-VN. Verbs like *up-regulate* (in the 'Change Activity' class) and *down-regulate* (in the 'Suppress' class) which are the direct indicators of their corresponding relation types in CHEMPROT, are available in our lexicon.

For many NLP applications, lexical classes are useful for their abilities to capture generalizations about a range of linguistic properties: by retrofitting word representations with our lexical resource, 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 corpora individually. Our lexical classes provide different levels of generalization power to support tasks of various needs, from the coarse-grained level of 16 classes to the fine-grained one of 50-classes. 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.

6.3 Chapter summary

In this chapter, we have introduced and evaluated an approach to facilitate cost-effective development of verb lexicons. In Section 6.1, we describe how our approach can automatically identify a set of useful contributive contexts for learning biomedical verb representations from large amounts of texts without manual feature engineering. Direct evaluation of the resulting models against Bio-SimVerb shows promising results when representation learning is performed using verb-related contexts. We further apply our verb-optimized representation models as features to induce a large-scale resource — a verb lexicon aims at describing verbs in the area of biomedicine. Human validation by linguists and biologists reveal that the lexicon, as induced using our optimized representation, is highly accurate and includes novel, valid member verbs and classes. Then, in Section 6.2, we evaluate our lexicon in the contexts of text classification and relation classification, it brings about a clear improvements over the raw verb features in most tasks, suggesting that the classification of 957 new verbs created by our approach (details in Section 6.1.4) and released with this thesis can be used to readily support application tasks in biomedicine.

Chapter 7

Conclusions

In this chapter, we summarize the contributions of this thesis and outline directions for future research.

7.1 Contributions of this thesis

The main contribution of this thesis is to advance the research of representation learning by improving its applicability across domains and word-types. Here we describe how the key six objectives of this work (given in Chapter 1 and summarized below) have been addressed.

1. Investigation of how the cutting-edge representation learning methods developed for general English can be transferred to the biomedical domain.
2. Development of an intrinsic evaluation dataset that can better reflect how individual representation models perform in extrinsic/downstream tasks.
3. Development of an intrinsic evaluation dataset that can measure the quality of verb representations in the biomedical domain.
4. Development of a technique that facilitates automatic lexical acquisition for biomedical verbs.
5. Reduction of the burden of feature engineering in automatic lexical acquisition.
6. Evaluation of the quality of our automatically-induced lexicon with domain experts, as well as its utility in supporting downstream tasks.

In Chapter 3, we conducted a large-scale experiment to find the best training settings for learning word representations from biomedical texts. These settings include corpora size,

model architectures and hyper-parameter values. Our experiments resulted in several key findings: First, we observed that a larger corpus does not necessarily guarantee better results in actual tasks. Also, we showed that optimization of hyper-parameters could significantly improve the performance of vectors. The optimized training configurations are highly domain-specific, and thus can serve as a reference for researchers who use neural word embeddings in biomedical NLP. More importantly, we find that one hyper-parameter (the context window size) leads to contradictory results between intrinsic and extrinsic evaluations.

In Chapter 4, we further investigated the intrinsic-extrinsic contradiction using general-domain datasets. We generated a set of word representations with varying context window sizes and compared their performance in intrinsic and extrinsic evaluations, showing that these evaluations yield mutually inconsistent results. Among all the benchmarks explored in our study, only SimLex-999 [Hill et al., 2015] proved to be a good predictor of downstream performance. We further considered the differentiation between relatedness (association) and similarity (synonymy) as an explanatory factor, noting that SimLex-999 stands out among other benchmark datasets in distinguishing highly similar concepts (*male, man*) from highly related but dissimilar ones (*computer, keyboard*).

In Chapter 5, we presented two new comprehensive resources targeting the evaluation of word representations in biomedicine. These resources, Bio-SimVerb and Bio-SimLex, address the intrinsic-extrinsic contradiction and can be used for evaluations of verb and noun representations respectively. In our experiments, we computed the Pearson's correlation between performances on intrinsic and extrinsic tasks using twelve popular state-of-the-art representation models (e.g. **word2vec** models). The intrinsic–extrinsic correlations using our datasets were notably higher than with previous intrinsic evaluation benchmarks such as UMNSRS and MayoSRS. Besides, when evaluating representation models for their abilities to capture verb and noun semantics individually, we showed a considerable variation between performances across all models. This result highlights not only the importance of developing dedicated evaluation resources for NLP in biomedicine for particular word-types (e.g. verbs), but also the need to identify the most accurate methods for learning type-specific representations.

In Chapter 6, we proposed and evaluated an automatic verb classification approach to facilitate cost-effective development of a verb lexicon. From the methodological point of view, our work constitutes the first effort on applying a state-of-the-art architecture for neural representation learning to biomedical verb classification. In terms of our contribution to representation learning, while previous works have shown that such neural models can be highly effective for learning linguistic properties from large corpora, there has been little work on fine-tuning the models for word-type specific tasks (e.g. verbs). Our work demonstrates

that the learning of verb-specific representation is highly context-sensitive. In particular, we identify the contexts that are essential for training representation for biomedical verbs. Our study can facilitate the development of different learning approaches for word-type specific representations as well as support researchers in biomedicine to better understand the syntactic and semantic properties of verbs in biomedical texts. On the other hand, as a verb classification method, our method is attractive in terms of avoiding the heavy feature engineering involved in most previous approaches. The human evaluation reveals that the lexicon, as induced created by our method, is highly accurate. Additionally, our promising results in the task-based evaluation empirically show that the resource can be used to support NLP applications in biomedicine.

7.2 Future directions

Here, we discuss how research reported in this thesis can be developed further in the future.

7.2.1 Hyper-parameter tuning

Regarding the optimization of training settings for representation models, we tuned individual parameters in isolation. As the next step, we can study the effect of tuning two or more parameters simultaneously. For example, the hyper-parameter *min-count* controls the minimum occurrences for a word to be included in representation learning. Careful tuning of this parameter both separately and jointly with associated parameters such as *samp* (which defines the occurrences for high-frequency words to be down-sampled) may offer further opportunities for improvement.

7.2.2 Evaluation resources

Regarding Bio-SimLex and Bio-SimVerb (our novel intrinsic evaluation datasets for biomedical representation), they are created to measure how well the notion of word similarity according to humans is captured by the word representations. Apart from word similarity, there are other semantic relations which are important for language understanding. One example is the 'hyponymy–hypernymy' (a.k.a 'is-a') relation that exists between concept groups such as *mammal* and their constituent members: *lion* or *tiger*. Being one of the essential linkages between entities as found in many biomedical ontologies (e.g. gene ontology), the 'is-a' relation underlines the lexical entailment relation. The ability to effectively model both lexical and phrasal entailment like humans can extend the usefulness of word representations to many related applications, such as question answering, information retrieval

and text summarization. For example, to answer a question such as 'Which insects can fly?', a question-answering system has to distinguish that a bee or a butterfly are types of insects, whereas an eagle or a pigeon are not. While intrinsic evaluation resources for lexical entailment have recently been developed for the general domain [Vulić et al., 2017a], there is a lack of similar resources in biomedicine, which suggests a potential research direction for this work.

We observe a positive correlation between the performance of representation models on Bio-SimLex and biomedical named entity recognition (NER). It is reasonable to expect that the evaluation of noun representations (Bio-SimLex) is more relevant to the performance of NER than evaluation of verb representations (Bio-SimVerb). In the future, it would be interesting to further assess the correlation between performance on Bio-SimVerb and other extrinsic tasks, such as relation typing, where performance is closely-related to the quality of verb representations.

7.2.3 Verb classification

We explore methods for fine-tuning existing neural representation methods for verb classification. This methodology can be improved in various directions: First, our representation models are trained on word co-occurrence frequencies to capture verb semantics on the word-level. Because many word formations in biomedicine follow regular patterns (e.g. *phosphorylate* and *dephosphorylate*), it might be possible to improve representation learning by incorporating both word and character-level information. In the future, we can explore other representation learning techniques for verb classification including FastText [Bojanowski et al., 2017] where learning procedure takes into account the morphological (subword) information.

Another potential research avenue is to improve the quality of representation learning through context modelling. Currently, we experiment with dependency-based contexts, showing that they can be useful in producing large semantically meaningful groups of classes. Nevertheless, there are a few cases where semantically dissimilar verbs are misclassified together because they share similar syntax. It indicates that there is room for improvement in identifying other discriminative contexts.

In this thesis, we used a supervised approach to verb classification (Nearest Centroid Classifier). While this provides an immediate benefit in terms of the accuracy of verbs classified, it requires a fixed set of pre-defined verb classes as part of the training data. To allow an unsupervised discovery of novel verb classes and subclasses, one idea for future work would be to improve the performance of unsupervised clustering algorithms with a small amount of supervision. It can take the form of labels on the data (seeds), constraints,

or user feedback. This type of approach, commonly known as semi-supervised clustering, not only can group candidates using the classes learned from the seed data, but it can also extend and modify the existing set of classes as needed to reflect other regularities in the data. Studies of this nature are emerging [Cuba Gyllensten and Sahlgren, 2018; Kuo et al., 2008] and it would be interesting to investigate how they can be applied to our task to reduce the need for pre-defined classes while maintaining promising precision.

In the last part of this thesis, we evaluated our automatically-created lexicon on text classification and relation classification. Our task-based evaluation can be further extended in various directions. First, we used retrofitting to incorporate the lexical features into word representations. This method has been widely-used in biomedicine and thus can serve as a baseline method for comparison. However, it focuses on synonyms within the same lexical class, and the 'hyponymy-hypernymy' relation between classes/sub-classes is largely neglected [1]. Because our lexicon is hierarchical, we can explore other more sophistical methods in the future. Examples included **LEAR** [Vulić, 2018], which utilizes the asymmetric relation of lexical entailment (the 'IS-A' relation). Second, retrofitting likes many other post-processing approaches of similar nature, are limited to updating only the vectors of words appearing in external lexicons (i.e. seen words), leaving the vectors of all other words unchanged (i.e. unseen words that only appear in either the models or the lexicons). In recent work of Ponti et al. [2018], they use the pre-retrofitted and post-retrofitted vector space to train a global specialization function (a.k.a. a transformer). This transformer will then be applied to the large subspace of unseen words to update their vectors. Their method effectively extends the specialization of word representation to the full vocabulary of the input distributional vector space, yet, it is carried out on general lexical resources with a mixed word-types (e.g. WordNet). To apply their method on lexicons of a specific word-type like our verb lexicon, a potential research avenue is to explore ways to guide the specialization to verb representations. Last, regarding our evaluation results, an in-depth performance/error analysis based on each classified-category can be carried out to investigate the effects of our lexical classes on individual document/relation types. Additionally, many other important NLP tasks and datasets can also be considered in the future.

Our current work constitutes the first effort on applying a state-of-the-art architecture for neural representation learning to biomedical verb classification. The preliminary experiment suggests that the verb resource, as created by our method, can be used to support application tasks in biomedicine. Our plan is to ultimately use our automatically-created lexicon to support the development of BioVerbNet via expert validation - an approach that can yield a fully accurate computational resource and enriched taxonomy with novel classes in biomedicine.

[1] In retrofitting, all sub-classes are merged and regarded as a unified classes

Once developed, such verb resource will provide a welcome addition to lexical resources in biomedicine which largely focus on nouns (e.g. the UMLS Metathesaurus mainly covers noun concepts) or a limited set of verbs (e.g. the BioLexicon provides the syntactic and semantic information of 168 verbs commonly used in E.Coli).

Viewed as a whole, we believe that the contributions presented by this thesis demonstrates the practical application of representation learning across domains (biomedicine) and word-types (verbs). In particular, our work shows that neural word embeddings can be used to benefit biomedical NLP in many ways.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of NAACL-HLT'09*, pages 19–27, 2009.

David J Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer, 1985.

Sophia Ananiadou and John Mcnaught. *Text mining for biology and biomedicine*. Artech House, Norwood, MA, USA, 2006.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics, 1998.

Simon Baker. *Semantic text classification for cancer text mining*. PhD thesis, University of Cambridge, 2018.

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

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

Simon Baker, Anna Korhonen, and Sampo Pyysalo. Cancer hallmark text classification using convolutional neural networks. *BioTxtM 2016*, page 1, 2016.

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

Mohit Bansal, Kevin Gimpel, and Karen Livescu. Tailoring continuous word representations for dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 809–815, 2014.

Libby Barak, Afsaneh Fazly, and Suzanne Stevenson. Learning verb classes in an incremental model. In *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics*, pages 37–45, 2014.

Marco Baroni and Alessandro Lenci. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10. Association for Computational Linguistics, 2011.

Colin Batchelor. Molecular process ontology, 2017. URL http://purl.obolibrary.org/obo/mop.owl.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, pages 1137–1155, 2003.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35 (8):1798–1828, 2013.

Steven Bird. Nltk: the natural language toolkit. In *Proceedings of COLING/ACL demos*, pages 69–72, 2006.

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

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

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, 2014.

Ryan R Brinkman, Mélanie Courtot, Dirk Derom, Jennifer M Fostel, Yongqun He, Phillip Lord, James Malone, Helen Parkinson, Bjoern Peters, Philippe Rocca-Serra, et al. Modeling biomedical experimental processes with obi. *Journal of biomedical semantics*, 1(1): S7, 2010.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4): 467–479, 1992.

Susan Windisch Brown, Dmitriy Dligach, and Martha Palmer. Verbnet class assignment as a wsd task. In *Proceedings of the Ninth International Conference on Computational Semantics*, pages 85–94. Association for Computational Linguistics, 2011.

Allen C Browne, Alexa T McCray, and Suresh Srinivasan. The specialist lexicon. *National Library of Medicine Technical Reports*, pages 18–21, 2000.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. Distributional semantics in technicolor. In *Proceedings of ACL'12*, pages 136–145, 2012.

Eugene Charniak and Mark Johnson. 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, 2005.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

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

Billy Chiu, Anna Korhonen, and Sampo Pyysalo. Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 1–6, 2016b.

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

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

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of ICML*, pages 160–167. ACM, 2008.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

BNC Consortium. The british national corpus, 2007. URL http://www.natcorp.ox.ac.uk.

UniProt Consortium. Uniprot: a hub for protein information. *Nucleic acids research*, 43(D1): D204–D212, 2014.

Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. A neural network multi-task learning approach to biomedical named entity recognition. *BMC bioinformatics*, 18 (1):368, 2017.

Amaru Cuba Gyllensten and Magnus Sahlgren. Distributional term set expansion. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resource Association, 2018. URL http://aclweb.org/anthology/L18-1405.

Marie-Catherine De Marneffe and Christopher D Manning. Stanford typed dependencies manual. Technical report, Technical report, Stanford University, 2008.

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

Rezarta Islamaj Doğan and Zhiyong Lu. An improved corpus of disease mentions in pubmed citations. In *Proceedings of the 2012 workshop on biomedical natural language processing*, pages 91–99. Association for Computational Linguistics, 2012.

Manaal Faruqui. Retrofitting, 2015. URL https://github.com/mfaruqui/retrofitting.

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

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. Problems with evaluation of word embeddings using word similarity tasks. In *Proc. of the 1st Workshop on Evaluating Vector Space Representations for NLP*, 2016. URL http://arxiv.org/pdf/1605.02276v1.pdf.

Javi Fernández, Yoan Gutiérrez, José M Gómez, and Patricio Martınez-Barco. Gplsi: Supervised sentiment analysis in twitter using skipgrams. In *Proceedings of SemEval*, pages 294–299, 2014.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *Proceedings of WWW'01*, pages 406–414, 2001.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 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, 2013.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. SimVerb-3500: A Large-Scale Evaluation Set of Verb Similarity. In *EMNLP*, 2016.

Georgios V Gkoutos, Eain CJ Green, Ann-Marie Mallon, John M Hancock, and Duncan Davidson. Using ontologies to describe mouse phenotypes. *Genome biology*, 6(1):R8, 2004.

Jennifer Golbeck, Gilberto Fragoso, Frank Hartel, Jim Hendler, Jim Oberthaler, and Bijan Parsia. The national cancer institute's thesaurus and ontology. *Web Semantics: Science, Services and Agents on the World Wide Web*, 1(1), 2011.

Rodrigo Rafael Villarreal Goulart, Vera Lúcia Strube de Lima, and Clarissa Castellã Xavier. A systematic review of named entity recognition in biomedical texts. *Journal of the Brazilian Computer Society*, 17(2):103–116, 2011.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. Revisiting embedding features for simple semi-supervised learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 110–120, 2014.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48, 2017.

Melissa A Haendel, Fabian Neuhaus, David Osumi-Sutherland, Paula M Mabee, Jos LV Mejino Jr, Chris J Mungall, and Barry Smith. Caro–the common anatomy reference ontology. In *Anatomy Ontologies for Bioinformatics*, pages 327–349. Springer, Berlin, 2008.

Kai Hakala, Suwisa Kaewphan, Tapio Salakoski, and Filip Ginter. Syntactic analyses and named entity recognition for pubmed and pubmed central—up-to-the-minute. *ACL 2016*, page 102, 2016.

Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. Large-scale learning of word relatedness with constraints. In *Proceedings of SIGKDD'12*, pages 1406–1414, 2012.

Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

Janna Hastings, Paula de Matos, Adriano Dekker, Marcus Ennis, Bhavana Harsha, Namrata Kale, Venkatesh Muthukrishnan, Gareth Owen, Steve Turner, Mark Williams, et al. The chebi reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic acids research*, 41(D1):D456–D463, 2013.

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

Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 2015.

Ray Jackendoff. *Semantic structures*, volume 18. MIT press, 1992.

Eric Joanis, Suzanne Stevenson, and David James. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(3):337–367, 2008.

Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001. URL http://www.scipy.org/. [Online; accessed <2017-03-01>].

Pentti Kanerva, Jan Kristoferson, and Anders Holst. Random indexing of text samples for latent semantic analysis. In *In Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 103–6. Erlbaum, 2000.

Daisuke Kawahara, Daniel Peterson, Octavian Popescu, and Martha Palmer. Inducing example-based semantic frames from a massive amount of verb uses. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 58–67, 2014a.

Daisuke Kawahara, Daniel W Peterson, and Martha Palmer. A step-wise usage-based method for inducing polysemy-aware verb classes. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1030–1040, 2014b.

Kendall and Maurice George. *Rank Correlation Methods, Second Edition, Revised*. Charles Griffin and Company Ltd, London, England, 1955.

Warren A Kibbe, Cesar Arze, Victor Felix, Elvira Mitraka, Evan Bolton, Gang Fu, Christopher J Mungall, Janos X Binder, James Malone, Drashtti Vasant, et al. Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic acids research*, page gku1011, 2014.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182, 2003.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of JNLPBA*, pages 70–75, 2004.

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

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of ICLR'15*, 2015.

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

Karin Kipper-Schuler. *VerbNet: a broad-coverage, comprehensive verb lexicon*. PhD thesis, Computer and Information Science Department, Universiy of Pennsylvania, Philadelphia, PA, 2005. URL http://repository.upenn.edu/dissertations/AAI3179808/.

Anna Korhonen, Yuval Krymolowski, and Nigel Collier. 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, 2006.

Anna Korhonen, Yuval Krymolowski, and Nigel Collier. The choice of features for classification of verbs in biomedical texts. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 449–456. Association for Computational Linguistics, 2008.

Aris Kosmopoulos, Ion Androutsopoulos, and Georgios Paliouras. Biomedical semantic indexing using dense word vectors in bioasq. *Journal Of Biomedical Semantics*, 2015.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1):S2, 2015.

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

Michael Krauthammer and Goran Nenadic. Term identification in the biomedical literature. *Journal of biomedical informatics*, 37(6):512–526, 2004.

Jin-Shea Kuo, Haizhou Li, and Ying-Kuei Yang. Active learning for constructing transliteration lexicons from the web. *Journal of the Association for Information Science and Technology*, 59(1):126–135, 2008.

Thomas K Landauer and Susan T Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

Gabriella Lapesa and Stefan Evert. A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, 2:531–545, 2014.

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

Angeliki Lazaridou, Eva Maria Vecchi, and Marco Baroni. Fish transporters and miracle homes: How compositional distributional semantics can help np parsing. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1908–1913, 2013.

Ben Lengerich, Andrew Maas, and Christopher Potts. 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, 2018. URL http://aclweb.org/anthology/C18-1205.

Ulf Leser and Jörg Hakenberg. What makes a gene name? named entity recognition in the biomedical literature. *Briefings in bioinformatics*, 6(4):357–369, 2005.

Beth Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago press, Chicago, IL 60637, USA, 1993.

Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proceedings of ACL'14*, pages 302–308, 2014.

Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

Jianguo Li and Chris Brew. Which are the best features for automatic verb classification. *Proceedings of ACL-08: HLT*, pages 434–442, 2008.

Dekang Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 768–774. Association for Computational Linguistics, 1998.

Dekang Lin, Kenneth Ward Church, Heng Ji, Satoshi Sekine, David Yarowsky, Shane Bergsma, Kailash Patil, Emily Pitler, Rachel Lathbury, Vikram Rao, et al. New tools for web-scale n-grams. In *LREC*. Citeseer, 2010.

Wang Ling, Chris Dyer, Alan W. Black, and Isabel Trancoso. Two/too simple adaptations of word2vec for syntax problems. In *HLT-NAACL*, 2015a.

Wang Ling, Yulia Tsvetkov, Silvio Amir, Ramon Fermandez, Chris Dyer, Alan W Black, Isabel Trancoso, and Chu-Cheng Lin. Not all contexts are created equal: Better word representations with variable attention. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1367–1372. Association for Computational Linguistics, 2015b. doi: 10.18653/v1/D15-1161. URL http://aclweb.org/anthology/D15-1161.

Thomas Lippincott, Laura Rimell, Karin Verspoor, and Anna Korhonen. Approaches to verb subcategorization for biomedicine. *Journal of biomedical informatics*, 46(2):212–227, 2013.

Carolyn E Lipscomb. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

Haibin Liu, Tom Christiansen, William A Baumgartner, and Karin Verspoor. Biolemmatizer: a lemmatization tool for morphological processing of biomedical text. *Journal of biomedical semantics*, 3(1):3, 2012.

Thang Luong, Richard Socher, and Christopher Manning. Better word representations with recursive neural networks for morphology. In *Proceedings of CoNLL*, pages 104–113, 2013.

Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany, August 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P16-1101.

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

Donna Maglott, Jim Ostell, Kim D Pruitt, and Tatiana Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic acids research*, 33(suppl_1):D54–D58, 2005.

Olga Majewska, Diana McCarthy, Ivan Vulić, and Anna Korhonen. Acquiring verb classes through bottom-up semantic verb clustering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2018.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119, 2013b.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, 2013c.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38 (11):39–41, 1995.

George A Miller and Walter G Charles. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28, 1991.

Andriy Mnih and Geoffrey E Hinton. A scalable hierarchical distributed language model. In *Advances in neural information processing systems*, pages 1081–1088, 2009.

Anupam Mondal, Dipankar Das, Erik Cambria, and Sivaji Bandyopadhyay. Wme 3.0: An enhanced and validated lexicon of medical concepts. In *Proceedings of the 9th Global WordNet Conference*. Global WordNet Association, 2017.

Luisa Montecchi-Palazzi, Ron Beavis, Pierre-Alain Binz, Robert J Chalkley, John Cottrell, David Creasy, Jim Shofstahl, Sean L Seymour, and John S Garavelli. The psi-mod community standard for representation of protein modification data. *Nature biotechnology*, 26(8):864–866, 2008.

Nikola Mrkšic, Diarmuid OSéaghdha, Blaise Thomson, Milica Gašic, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. Counter-fitting Word Vectors to Linguistic Constraints. In *Proceedings of NAACL-HLT*, pages 142–148, 2016.

TH Muneeb, Sunil Kumar Sahu, and Ashish Anand. Evaluating distributed word representations for capturing semantics of biomedical concepts. In *Proceedings of ACL-IJCNLP*, page 158, 2015.

Christopher J Mungall, Carlo Torniai, Georgios V Gkoutos, Suzanna E Lewis, and Melissa A Haendel. Uberon, an integrative multi-species anatomy ontology. *Genome biology*, 13(1): R5, 2012.

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407, 2004.

Nhung TH Nguyen, Makoto Miwa, Yoshimasa Tsuruoka, Takashi Chikayama, and Satoshi Tojo. Wide-coverage relation extraction from medline using deep syntax. *BMC bioinformatics*, 16(1):107, 2015.

U.S. NLM. Broad subject terms, 2017. URL https://wwwcf.nlm.nih.gov/serials/journals/index.cfm.

Diarmuid Ó Séaghdha and Ann Copestake. Semantic classification with distributional kernels. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 649–656. Association for Computational Linguistics, 2008.

Tomoko Ohta, Sampo Pyysalo, Jun'ichi Tsujii, and Sophia Ananiadou. Open-domain anatomical entity mention detection. In *Proceedings of the workshop on detecting structure in scholarly discourse*, pages 27–36. Association for Computational Linguistics, 2012.

Naoaki Okazaki and Sophia Ananiadou. Building an abbreviation dictionary using a term recognition approach. *Bioinformatics*, 22(24):3089–3095, 2006.

Sebastian Padó and Mirella Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, 2007.

Serguei Pakhomov, Bridget McInnes, Terrence Adam, Ying Liu, Ted Pedersen, and Genevieve B Melton. Semantic similarity and relatedness between clinical terms: an experimental study. In *Proceedings of AMIA*, volume 2010, page 572, 2010.

Serguei VS Pakhomov, Ted Pedersen, Bridget McInnes, Genevieve B Melton, Alexander Ruggieri, and Christopher G Chute. Towards a framework for developing semantic relatedness reference standards. *Journal of biomedical informatics*, 44(2):251–265, 2011.

Judea Pearl. *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1984. ISBN 0-201-05594-5.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, volume 14, pages 1532–1543, 2014.

Daniel Peterson, Jordan Boyd-Graber, Martha Palmer, and Daisuke Kawahara. Leveraging verbnet to build corpus-specific verb clusters. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 102–107, 2016.

Steven Pinker. *Learnability and cognition: The acquisition of argument structure*. MIT press, 1989.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 282–293. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/D18-1026.

Sampo Pyysalo and Sophia Ananiadou. Anatomical entity mention recognition at literature scale. *Bioinformatics*, page btt580, 2013.

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

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

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of WWW'11*, pages 337–346, 2011.

Marek Rei, Gamal K. O. Crichton, and Sampo Pyysalo. Attending to characters in neural sequence labeling models. In *COLING*, 2016.

Laura Rimell, Thomas Lippincott, Karin Verspoor, Helen L Johnson, and Anna Korhonen. Acquisition and evaluation of verb subcategorization resources for biomedicine. *Journal of biomedical informatics*, 46(2):228–237, 2013.

Will Roberts and Markus Egg. A comparison of selectional preference models for automatic verb classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 511–522, 2014.

Cornelius Rosse and José LV Mejino Jr. The foundational model of anatomy ontology. In *Anatomy Ontologies for Bioinformatics*, pages 59–117. Springer, Berlin, 2008.

Herbert Rubenstein and John B Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

Rune Sætre, Kazuhiro Yoshida, Akane Yakushiji, Yusuke Miyao, Yuichiro Matsubayashi, and Tomoko Ohta. Akane system: protein-protein interaction pairs in biocreative2 challenge, ppi-ips subtask. In *Proceedings of BioCreative II*, pages 209–212, 2007.

Dage Särg. Hierarchical clustering of estonian verb constructions. *ESSLLI 2017 Student Session*, page 221, 2017.

Carolina Scarton, Lin Sun, Karin Kipper-Schuler, Magali Sanches Duran, Martha Palmer, and Anna Korhonen. Verb clustering for brazilian portuguese. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 25–39. Springer, 2014.

Michael Schmitz, Robert Bart, Stephen Soderland, Oren Etzioni, 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*, pages 523–534. Association for Computational Linguistics, 2012.

Roy Schwartz, Roi Reichart, and Ari Rappoport. Symmetric pattern based word embeddings for improved word similarity prediction. In *CoNLL*, volume 2015, pages 258–267, 2015.

Erik Segerdell, Jeff B Bowes, Nicolas Pollet, and Peter D Vize. An ontology for xenopus anatomy and development. *BMC Developmental Biology*, 8(1):92, 2008.

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

Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, et al. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255, 2007.

Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. Overview of biocreative ii gene mention recognition. *Genome biology*, 9(Suppl 2): 1–19, 2008.

Pontus Stenetorp, Hubert Soyer, Sampo Pyysalo, Sophia Ananiadou, and Takashi Chikayama. Size (and domain) matters: Evaluating semantic word space representations for biomedical text. In *Proceedings of SMBM*, 2012.

Lin Sun. *Automatic induction of verb classes using clustering*. PhD thesis, University of Cambridge, 2013.

Lin Sun, Anna Korhonen, and Yuval Krymolowski. Automatic classification of english verbs using rich syntactic features. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*, 2008a.

Lin Sun, Anna Korhonen, and Yuval Krymolowski. Verb class discovery from rich syntactic data. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 16–27. Springer, 2008b.

He Tan. 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)*, pages 46–53, 2014.

Lorraine Tanabe, Natalie Xie, Lynne H Thom, Wayne Matten, and W John Wilbur. Genetag: a tagged corpus for gene/protein named entity recognition. *BMC bioinformatics*, 6(1):S3, 2005.

Buzhou Tang, Hongxin Cao, Xiaolong Wang, Qingcai Chen, and Hua Xu. Evaluating word representation features in biomedical named entity recognition tasks. *BioMed research international*, 2014, 2014.

Paul Thompson, John McNaught, Simonetta Montemagni, Nicoletta Calzolari, Riccardo Del Gratta, Vivian Lee, Simone Marchi, Monica Monachini, Piotr Pezik, Valeria Quochi, et al. The biolexicon: a large-scale terminological resource for biomedical text mining. *BMC bioinformatics*, 12(1):397, 2011.

Erik F Tjong Kim Sang and Sabine Buchholz. Introduction to the conll-2000 shared task: Chunking. In *Proceedings of CoNLL'00*, pages 127–132, 2000.

Erik F Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL'03*, pages 142–147, 2003.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. Evaluation of word vector representations by subspace alignment. In *Proc. of EMNLP*, 2015.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics, 2010.

Peter D Turney. Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research*, pages 533–585, 2012.

Ceri E Van Slyke, Yvonne M Bradford, Monte Westerfield, and Melissa A Haendel. The zebrafish anatomy and stage ontologies: representing the anatomy and development of danio rerio. *Journal of biomedical semantics*, 5(1):12, 2014.

Giulia Venturi, Simonetta Montemagni, Simone Marchi, Yutaka Sasaki, Paul Thompson, John McNaught, and Sophia Ananiadou. Bootstrapping a verb lexicon for biomedical information extraction. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 137–148. Springer, 2009.

Andreas Vlachos, Anna Korhonen, and Zoubin Ghahramani. Unsupervised and constrained dirichlet process mixture models for verb clustering. In *Proceedings of the workshop on geometrical models of natural language semantics*, pages 74–82. Association for Computational Linguistics, 2009.

Ivan Vulić. Injecting lexical contrast into word vectors by guiding vector space specialisation. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 137–143. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/W18-3018.

Ivan Vulić and Anna Korhonen. Is "universal syntax" universally useful for learning distributed word representations? In *Proceedings of ACL*, pages 518–524, 2016.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. Hyperlex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics*, 43(4): 781–835, 2017a.

Ivan Vulić, Nikola Mrkšić, and Anna Korhonen. Cross-lingual induction and transfer of verb classes based on word vector space specialisation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2548, 2017b.

Ivan Vulić, Roy Schwartz, Ari Rappoport, Roi Reichart, and Anna Korhonen. Automatic selection of context configurations for improved class-specific word representations. In *Proceedings of CoNLL*, pages 112–122, 2017.

Ramona L Walls, Balaji Athreya, Laurel Cooper, Justin Elser, Maria A Gandolfo, Pankaj Jaiswal, Christopher J Mungall, Justin Preece, Stefan Rensing, Barry Smith, et al. Ontologies as integrative tools for plant science. *American journal of botany*, 99(8):1263–1275, 2012.

Tuangthong Wattarujeekrit, Parantu K Shah, and Nigel Collier. Pasbio: predicate-argument structures for event extraction in molecular biology. *BMC bioinformatics*, 5(1):155, 2004.

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

Aaron Steven White, Rachel Dudley, Valentine Hacquard, and Jeffrey Lidz. Discovering classes of attitude verbs using subcategorization frame distributions. In *Proceedings of the 43rd Meeting of the North East Linguistic Society*, volume 43, 2014.

John Wieting, Mohit Bansal, Kevin Gimpel, Karen Livescu, and Dan Roth. From Paraphrase Database to Compositional Paraphrase Model and back. *Transactions of the Association for Computational Linguistics*, 3:345–358, 2015.

Wikipedia. Wikipedia, the free encyclopedia, 2016. https://dumps.wikimedia.org/enwiki/latest/.

Dongqiang Yang and David MW Powers. Verb similarity on the taxonomy of wordnet. In *Proceedings of GWC'06*, 2006.

Mo Yu and Mark Dredze. 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, 2014.

Zhiguo Yu, Trevor Cohen, Byron Wallace, Elmer Bernstam, and Todd Johnson. 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, 2016.

Appendix A

Guidelines for classification of biomedical verbs for an automatically-created resource

A.1 Background

The experiment aims to extend the small biomedical verb classification of Korhonen et al. [2006] with the view facilitating the creation of an automatically created resource. The small classification contains 192 verbs organized into a 3-level taxonomy consisting of 16, 34 and 50 classes. We have now applied an automatic classification approach (described in Chapter 6) to create an extended classification. It consists of 1,149 verbs in total (the 192 original ones plus 957 new ones) that have been grouped into the original class taxonomy based on their shared meanings and syntax according to our learning technique.

Your task is to verify whether these new candidate verbs are really similar in terms of their meanings as well as syntactic patterns to existing verbs in the original classification. Here is our initial proposal for how the task could be conducted.

The task has the following the steps (in blue, tasks to be answered in the Excel spreadsheet: **Answer.xlsx**):

A.2 Task A: Decide whether new verbs in each verb class share the similar meanings and syntactic patterns

A.2.1 Materials

You will be provided with 3 documents to support this task. They are:

1. **Question.xlsx**: The list of verbs grouped into classes, with descriptions of each class.
2. **Answer.xlsx**: An Excel spreadsheet for recording the updated index of the class for each verb based on your perception.
3. **Examples** (folder): Example sentences for each verb.

Please download and unzip the materials from:
https://drive.google.com/open?id=1tbV92X7fG13ereSIxo2oTC0Ly1zGbXqM

A.2.2 Task description

Open the file: **Question.xlsx**, you will see verbs grouped into classes based on their shared meanings and syntax. They are organised in five columns (see Figure A.1) as follows:

Class Name: The name of each class.

Sub-class Name: The name of each sub-class.

Class index: The unique identifier (which you will need to use throughout the entire task).

Example Verbs: Example verbs for each class from the original 192 classification.

New candidates: The list of new candidate verbs for verification.

Your task is to decide whether each new candidate verb (i.e. *New Candidates* in Fig A.1) has been assigned to the right class/sub-class based on your interpretation of the *Example verbs* in each class, as well as the sentence examples we provided for each verb (in the **Example** folder, as describes in Section A.2.2). You should give your answers on the file we provided (**Answer.xlsx**, as describes in Section A.2.2).

A	B	C	D	E
Class Name	**Sub-class Name**	**Class Index**	**Example Verbs**	**New candidates**
Biochemical events	Express	2.1.0	express overexpress	coexpress,secrete
	Biochemical modification	2.2.1	dephosphorylate phosphorylate	deacetylate,ubiquitinate,acetylate
	Cleave	2.2.2	cleave	
	Interact	2.3.0	react interact interfere cooperate colocalize coincide correlate	cosegregate,compensate,collaborate,participate,coevolve,correspond

.

Fig. A.1 A screen-shot of the subset of verb class in *Question.xlsx*. *Class Name* is the name of the top-level class. *Sub-class Name* is the name of each sub-class. *Class index* is the unique identifier of each class/sub-class. *Example Verbs* has the member verbs of each sub-class. *New candidates* contains verbs to be verified by annotators. They are separated from *Example verbs* by red line for distinction

Sentence examples

To help you understand how a verb is used in biomedical text, we provide about thirty example sentences from the corpus we used in our experiment, which illustrate the most common syntactic structures of each verb (in descending order, most common on top and least common at bottom). They are stored in folder: **Example** with the test verb as the filename. They are organized in 3 columns: The first column is the name of the dependency pattern exemplified in the sentence. The second column is the sentence example. The third column is the word in sentences corresponding to the syntactic pattern (see Figure A.2).

increase.txt

```
1  obj This seemed to increase illness-related strain and a need for defensive actions .   obj@strain
2  obj Disruption of yqhC offers a useful approach to increase furfural tolerance in bacteria .    obj@tolerance
3  obj Breast feeding does not appear to increase the risk of postpartum relapses .    obj@risk
4
5  subj#obj    Inappropriate use of emergency care services can increase hospital readmissions and related costs . subj@use obj@readmissions
6  subj#obj    Simulated altitude did not increase incidence of ECS .  subj@altitude obj@incidence
7  subj#obj    However , pharmacological PPARdelta activation did not increase T-cadherin expression . subj@activation obj@expression
8
```

Fig. A.2 A screen-shot of example sentences of *increase* (in Folder: *Example*). The first column contains common syntactic patterns for *increase* in descending order (e.g. *obj*=object). The second column stores the sentence example for using the corresponding pattern. The third column stores the corresponding words in the sentence for the pattern (e.g. *strain*)

Look into the sentence examples of each *New candidates* and *Example Verbs* in each class (as mentioned in Section A.2.2), decide if each new candidate verb has been assigned to the right class. Give your answers on our answer template in the pre-defined format, which is described in the next section.

New candidates	Current Class	Final Class
demethylate	1.1.1	
biotinylate	8.1.0	

Fig. A.3 A screen-shot of the answer sheet for annotators (filename: *Answer.xlsx*). *New candidates* contains verbs to be verified by annotators. *Current Class* is the index of the class where a verb currently assigned to. *Final Class* records the updated class indexes for the verbs after verified by annotators.

Answers template

Open the file: **Answer.xlsx**, you will see all the new candidates (Column 1) and the classes they are currently assigned to (Column 2). Please write down the *Class Index* (reference from *Question.xlsx*) you think each verb should be assigned to. Here, we use *demethylate* and *biotinylate* as examples (see Fig A.3), they are currently assigned to the Class *1.1.1* and Class *8.1.0* correspondingly. There are three options to choose:

1. If you think the verbs are correctly assigned, just put down the same class indexes as their suggested ones (i.e. *1.1.1* and *8.1.0*) in their corresponding cell in the *Final Class* column.

2. In contrast, If you think the verbs are incorrectly assigned:

   (a) If the mis-assigned verb should be in another class, please put down the corresponding class index. For example, if you think *demethylate* should be in the class: *Biochemical modification* (see Fig A.1), then put down its index: *2.2.1* in the *Final Class* column.

   (b) If at least two mis-assigned verbs can be part of an entirely new top class, please put down a new class index in the format: (*N+1.0.0*) where *N* is the current largest top-class index (By default, we have 16 top-level classes (*N=16*), so new index begins with *17.0.0*). For example, if you think *demethylate* and *biotinylate* can be part of an entirely new classes, and this is the first class index you create, then put down *17.0.0* in both of their cells in the *Final Class* column. Subsequent new class index will then be *18.0.0*, *19.0.0*...etc).

   (c) Any verbs you cannot find a good class for, please put in 0 as its class index in the *Final Class* column.

Give a final class index to each new candidate verb. HOWEVER, A VERB CAN ONLY BE ASSIGNED TO ONE CLASS/SUBCLASS ONLY!!!

Appendix B

An incidence matrix showing the class reassignments of verbs in our automatically-created lexicon

The new verbs judged as not valid were marked as candidates for reassignment to another existing class, or as members of a subclass or a new class altogether. An incidence matrix showing the class reassignments is presented in Table B.1. For instance, *exacerbate, aggravate* and *magnify*, found in the 'Inactivate' class, were highlighted as forming a separate cluster of similar verbs, while the verb *deacylate* found in the 'Release' class was reassigned to the 'Modify' class. In the general scientific domain, an example of reassignment involved verbs *display* and *exhibit*, found in the 'Encompass' class but considered better suited for the 'Indicate' class, within which four other candidates, *underline, underscore, highlight, emphasize*, were marked as forming a subclass of 'underline'-type verbs. Such cases demonstrate the potential of the classification method for also discovering valid novel classes not in the original classification.

	7.1.0 COLLECT	9.1.1 EXAMINE	9.3.0 INDICATE	10.1.3 CONDUCT	13.1.0 ENCOMPASS	14.0.0 CALL	16.0.0 APPEAR	1.1.2 SUPRESS	1.1.4 INACTIVATE	1.4.0 MODIFY	2.3.0 INTERACT	4.1.3 LABEL	8.3.1 TRANSPORT	11.0.0 RELEASE	NEW CLASS	**Counts**
7.1.0 COLLECT													transfuse			1
9.1.1 EXAMINE																0
9.3.0 INDICATE																0
10.1.3 CONDUCT																0
13.1.0 ENCOMPASS			display, exhibit													2
14.0.0 CALL																0
16.0.0 APPEAR																0
1.1.2 SUPRESS																0
1.1.4 INACTIVATE															exacerbate, aggravate, magnify	3
1.4.0 MODIFY															detoxify, metabolize	2
2.3.0 INTERACT																0
4.1.3 LABEL																0
8.3.1 TRANSPORT																0
11.0.0 RELEASE										deacylate						1

Table B.1 An incidence matrix showing the class reassignments of verbs in our automatically-created lexicon. It shows how verbs are reassigned from their original classes (rows) to their final classes (columns) as determined by human annotators. **Counts** refer to the total numbers of reassignments of each class. If the annotators cannot find a suitable class to fit-in a verb, it will be assigned to **New Class**.