A Continual Machine Learning Framework for Accelerating Scientific Discovery

---

A

Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

by

Kishlay Jha

August 2022

Acknowledgements

I would like to thank many people who have helped me along my path of pursuing a Ph.D. degree in computer science. Firstly, I would like to thank my advisor Professor Aidong Zhang who gave me this opportunity to purse PhD in her research lab. I am very grateful to have her as my mentor and consider myself extremely fortunate. I would like to thank my committee members Professor Hongning Wang, Professor Jundong Li, Professor Nathan Sheffield and Professor Jing Gao for their valuable suggestions for my research and for shaping my dissertation. I also appreciate the support and help from my collaborators, lab mates and friends. I would like to thank Guangxu Xun, Yaqing Wang, Vishrawas Gopalakrishnan, Fenglong Ma,Ye Yuan, Mengdi Huai, Jianhui Sun, Yaliang Li, Houping Xiao, Hongfei Xue, Qiuling Suo, and many more whose names are not included here. I would like to thank my parents, Raman Jha and Nootan Jha, for raising me to value education. I would like to thank my brother, Bhaskar Jha, for his support and encouragement. I would also like to thank my wife, Aditi Jha, for always being supportive of my decisions.

Abstract

Machine learning models, which have found tremendous success in several commercial applications where large-scale data is available (e.g., computer vision and natural language processing), are beginning to play an important role in scientific disciplines such as biomedicine. Over the past few years, several domain-specific knowledge discovery frameworks have been proposed. Despite significant advances made, current research trends in machine learning-based approaches have not kept pace, for two main reasons. 1) The rapid proliferation of biomedical literature (on average around 3,000 articles are published every day) necessitates the development of innovative systems that can continually acquire and adapt to the new data. However, the existing approaches usually adopt a static learning paradigm and thus are unable to handle this setting. 2) Since the existing approaches mainly assume a static setting, they do not factor in the temporal evolution of biomedical concepts. This is limiting because the biomedical concepts are known to periodically acquire new semantic sense and lose old ones. To address these aforementioned challenges, we propose to shift the research direction from the currently dominant paradigm of static learning to continual learning, wherein the proposed approach is able to transfer useful knowledge over time and process the newly available articles in an efficient yet accurate manner. Specifically, the proposed approach exploits the unique capabilities of self-supervised learning, supervised learning and lifelong learning to design a continual learning framework that progressively acquires new scientific knowledge, models the semantic evolution of biomedical concepts, and generates actionable insights (novel meaningful associations) that can drive new research frontiers.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

The constant influx of scientific articles and their easy accessibility via the World Wide Web (WWW) has made biomedical informatics a fast growing field [3]. Researchers in the field have thrived to make sense of huge number of academic publications, discovery notes, electronic medical records and other text materials leading to advancements of practical significance [4]. While this swift availability of scientific information has acted as an impetus for pacing research innovation, it has also overwhelmed researchers trying to survey published studies and construct new ideas. For instance, consider a novice researcher attempting to formulate a new hypothesis for the cures of *Diabetes*. In doing so, one might have to survey tens of thousands of existing publications (more than 400,000 in PubMed [3] alone) already written on *Diabetes*. This overloaded amount of information creates a fundamental bottleneck to scientific productivity, as it is almost impossible for one to process and analyze such a large volume of available material. To mitigate these issues, there has been a growing research interest among data/text mining researchers to develop computational models that are able to assist biomedical experts in forging analytically probable, medically sensible hypothesis for possible *in-vitro* clinical trials. Towards this end, Hypothesis generation (HG), a sub-problem of biomedical text-mining, aims to discover hitherto unknown connections by chaining together the already known and established scientific facts remaining dispersed across the corpus. Simply put, given an input concept of interest (e.g., disease or gene), HG attempts to find implicit connections (e.g., potential drug target or novel indicator of disease's

mechanism) that link them in a previously unknown but semantically meaningful way. As an illustration, consider the example of *Raynaud's disease* and *Fish Oils* markedly discussed in the literature [5, 6, 7]. Prior to 1985, there was no direct connection known between *Raynaud's disease* and *Fish Oils*. However, in 1986, after manually inspecting the titles of articles on both topics separately, researchers inferred (later clinically validated [8]) an association between them. Finding such meaningful implicit links is the essence of the problem that this proposal attempts to address.

In the past few decades, numerous studies based on distributional approaches [6, 9], graph-based methods [10, 11], and supervised machine learning [12] have been conducted to tackle this problem. However, these studies posses a few inherent drawbacks. First, a majority of these preceding approaches rely on a pre-defined structure (e.g., graph) and hence possibly risk missing *surprising* links that are not included in their route. Second, almost all of these studies assume that the domain is static. This is limiting because it is known that the biomedical domain is a highly evolving field with new facts being added evey now and then [13]. Meanwhile, some of the contemporary studies [10] have also attempted to use the triplets (subject-relation-object) obtained from SemRep [14] as their unit of analysis to perform hypothesis generation. Although promising, at present, the overall recall of predications extracted by SemRep is relatively low (55%) [15]. This might cause a substantial number of semantic associations between entities to be missed, resulting in inaccurate hypotheses.

Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) has presented us with a new capability to scour the world's scientific literature using the state-of-the-art Natural Language Processing (NLP) approaches. Amongst others, word embedding [16, 17] techniques, that are primarily based on multi-layer neural networks have the ability to parse the wealth of readily available textual information and identify underlying implicit connections. These embedding techniques learn continuous low dimensional vectors of words (commonly known as word embeddings) in a completely unsupervised manner. As these word embeddings have been shown to encode the *implicit* semantics at a granular level, they are well suited for the current task of interest. This is because HG itself can be thought of as identifying *implicit* connection across previously disjoint terms in the latent space. While there exist word embedding models such as Word2Vec [18] and GLoVe [17] that can generate vector representation of words from natural language text, these models assume a static world. This becomes problematic for task such as hypothesis generation where it is crucial to factor in the temporal

dynamics of medical concepts in order to generate accurate hypothesis. To overcome this challenge, we propose our solution to identify the implicit links by learning the subtle cues manifested in temporal association formation process. In other words, we aim to gain a holistic understanding of evolutionary association formation process, wherein the terms with a potential of forming a connection iteratively come closer to each other in each time-stamp. Generally speaking, our proposed framework automatically discovers new knowledge based on what it perceives as significant historical trend and causes for relationship formation. Powered by recent NLP techniques such as word-embeddings, this proposal provides a scalable approach to identify high quality novel postulates that could be of potential clinical interest to the biomedical researchers. Furthermore, the interpretability part of this research provides a systematic mechanism through which one can study the evolution of the association formation process over time slices. This form of explainability augments the output of hypothesis generation module and provides the researchers with necessary evidence required to validate the hypotheses.

1.2 Problem Statement

Although, related to the problem of link prediction [19, 20, 21], hypotheses generation differs markedly from it in the sense that the output or the hypotheses is backed with explainable evidence. Furthermore, it draws similarity to the deep QA systems which are widely studied in area of Information Retrieval. This similarity stems from the fact that the input to the hypotheses generation system could be formulated as a question: "Is *Fish oils* and *Raynaud's Disease* connected?". Questions like this are generally categorized as 'closed discovery' wherein one is interested in evidence that connects the two medical concepts in the question. A variation of this question is "What are the therapeutic options for *Raynaud's Disease*?". This type of questions is essentially a generalization of the "closed discovery" and requires efficient algorithms to search amongst all different possible answers. In other words, hypotheses generation can be thought to be as one of making connections across previously unconnected terms. Towards that goal, we model our solution to handle the process of the connection formation as one driven by temporal characteristics of the medical concepts. In other words, we understand evolutionary association formation process of, wherein the terms with a potential of forming a connection iteratively come closer to each other in each time-stamp. However, modeling such a behavior requires the capability to quantify the semantic

similarity between the terms and the changes between them over a time period. Apart from this, it opens up a new challenge of determining candidate connecting terms as focusing on all the medical concept would not only affect the efficiency of our model but also introduce noise in our output.

The input to the system is a pair of medical terms (a medical term and a meta-information in the case of open discovery). Along with these terms, a year is also provided that acts as a threshold and limits the framework to base its analyses only on the documents published before that date. This is an optional field. The job of the "*Hypotheses Generation Module*" is to list a series of postulates that relate the two input terms through intermediaries/connecting term, e.g., Fish oils → Beta-Thromboglobulin → Raynaud Disease. The "Ranking Module" is then responsible for determining which of these candidate connecting terms have a higher chance of materializing in the future. In other words, based on the temporal/evolutionary properties of these connecting terms, the model measures the likelihood of an edge forming between these terms and the input query terms. Since, the final output would be ranked list of connecting terms, this module is called the "Ranking Module". As the core task, i.e., finding the connecting terms, is cascaded to the previous step, one can treat this as a black-box, where various algorithms/techniques could be applied as a plug and play.

In this work, we will be discussing our approaches towards solving these two tasks. Namely, we exploit recent advances in the area of word-embeddings to automatically learn which of the hypotheses have the potential to be ranked higher by a ranking module; thus, allowing us concentrate only on high quality viable hypotheses. Detailed discussion on the approach is presented in [22]. We then extend the concept of word-embeddings and its application to include temporal aspect to measure novelty and importance of the hypotheses. Through this, we rank the generated hypotheses [23]. In the following sections, we briefly describe the concept of word-embeddings and then its application to efficient enumeration of connecting terms through a process we call 'self-learning'. These discussions are then followed by introduction incorporating temporal aspect to the word-embeddings and its usage in measuring the importance of the connecting terms.

1.3 Contributions and Dissertation Organization

This research introduces a novel framework to tackle the challenging issues raised by the unprecedented proliferation of massive biomedical data - both structured and unstructured. By advancing the deep learning powered NLP techniques, the proposed approach elucidates the capability to discern plausible associations that would otherwise elude the domain experts. The inter-disciplinary nature of our problem brings biomedical researchers and computer scientists to work together in a fresh and integrated view. This collective endeavor will expedite the discoveries in complex biomedical domain, by identifying latent associations between disjoint concepts and inferring new meanings behind the ones already identified. More specifically, the proposed framework learns the association formation process between concepts, by tightly incorporating their evolutionary features and semantic relations present in human curated knowledge-bases such as ontologies and lexicons. The innovative scientific aspects of this research include:

- The capability of proposed approach to model the gradual semantic evolution of medical concepts over time. This allows the biomedical practitioners to track and visualize the evolutionary trajectories of various concepts, thereby, achieving fine interpretability of the generated output.
- The intrinsic ability of proposed framework to learn subtle cues manifested in temporal drift enables it to reveal promising implicit relationships, which can be translated into explicit real-world connections.
- The "plug-and-play" nature of various research components provides opportunity for biomedical practitioners to seamlessly integrate our modules with their off-the-shelf ML algorithms, benefiting a plethora of biomedical applications.
- The proposed framework provides an ideal setting where the navigation of the hypothesis generation is personalized at a query level, allowing the practitioners to drill down onto aspects relevant to them and their specific analysis goals.
- The introduced framework is versatile enough to enable the users/biomedical practitioners run custom queries and analyses on their proprietary data, and at the same time allow them to leverage the knowledge from existing public repository knowledge bases (KBs).
- The inherent interpretability of modules assists us in gathering additional evidence to strengthen our results and further utilize it for experimental validation.

Apart from these we are also planning to release a suite of tools that can be readily employed on the datasets for knowledge discovery. Altogether, our main contribution lies in accelerating the scientific progress by integrating large-scale scientific biomedical corpus, identifying the implicit links that are relevant to a given query, and from these links suggesting hypotheses that are new, insightful, testable and likely to be true. The proposed research will advance the state-of-the-art in biomedical sciences by providing new, efficient, and powerful tools for research and analysis. In addition to comparing against a range of baselines and benchmark algorithms, we shall apply this framework in several kind of diseases. This will demonstrate the capability of proposed system to discover the reasons and mechanistic interactions in a variety of case studies. By including comprehensive biomedical literature and knowledge bases that are publicly available, we can perform collective discovery and analyses of the same, giving an insight of causal relationships and their evolution over time.

The organization of this dissertation is summarized as follows:

In chapter 2, we propose to develop an enhanced word embedding representation that jointly exploits both contextual information and available explicit biomedical knowledge to learn a high-quality word embeddings representation. Unlike existing approaches, the proposed methodology is more dexterous in its ability to handle relationships between indirectly related concepts. Furthermore, we propose a dynamic word embeddings model that is capable of modeling the temporal information of concepts present in diachronic biomedical corpus.

In chapter 3, we propose to learn temporally aware vector representation of medical concepts from the time-stamped text data, and in doing so provide a systematic approach to formalize the problem. More specifically, a dynamic word embedding based model that jointly learns the temporal characteristics of medical concepts and performs across time-alignment is proposed. Apart from capturing the evolutionary characteristics in an optimal manner, the model also factors in the implicit medical properties useful for a variety of bio-medical applications. Empirical studies conducted on two important bio-medical use cases validates the effectiveness of the proposed approach and suggests that the model not only learns quality embeddings but also facilitates intuitive trajectory visualizations.

In chapter 4, we propose to develop a novel model to be used for interpreting word

embeddings representations, that is capable of transforming any pre-trained word embeddings to a new space such that the hidden conceptual meaning of individual dimensions are revealed. To the best of our knowledge, we are among the first to study the interpretability of word embedding in the biomedical domain. By leveraging upon the principles of dictionary learning and exploiting the categorical knowledge present in the biomedical domain, the proposed model is capable of generating an interpretable word representation that resembles closely to the human-level intuition.

In chapter 5, we propose a novel representation learning approach that accurately preserves the intricate bipartite structure, and efficiently updates the node representations. Specifically, we design a customized autoencoder that captures the proximity relationship between nodes participating in the bipartite bicliques (2×2 sub-graph), while preserving both the global and local structures. Moreover, the proposed structure-preserving technique is carefully interleaved with the central tenets of continual machine learning to design an incremental learning strategy that updates the node representations in an online manner. Taken together, the proposed approach produces meaningful representations with high fidelity and computational efficiency. Extensive experiments conducted on several biomedical bipartite networks validate the effectiveness and rationality of the proposed approach.

In chapter 6, we propose a new representation learning approach that efficiently adapts the concept representations to the newly available data. Specifically, the proposed approach develops a knowledge-guided continual learning strategy wherein the accurate/stable context-information present in human-curated knowledge-bases is exploited to continually identify and retrain the representations of those concepts whose corpus-based context evolved coherently over time. Different from previous studies that mainly leverage the curated knowledge to improve the accuracy of embedding models, the proposed research explores the usefulness of semantic knowledge from the perspective of accelerating the training efficiency of embedding models. Comprehensive experiments under various efficiency constraints demonstrate that the proposed approach significantly improves the computational performance of biomedical word embedding models.

In chapter 6, we propose a new representation learning approach that progressively

fuses the semantic information from multiple KBs into the pretrained biomedical language models. Since most of the KBs in the biomedical domain are expressed as parent-child hierarchies, we choose to model the hierarchical KBs and propose a new knowledge modeling strategy that encodes their topological properties at a granular level. Moreover, the proposed continual learning technique efficiently updates the concepts representations to accommodate the new knowledge whilst preserving the memory efficiency of contextualized language models. Altogether, the proposed approach generates knowledge-powered embeddings with high fidelity and learning efficiency. Extensive experiments conducted on bioNLP tasks validate the efficacy of the proposed approach and demonstrates its capability in generating robust concept representations.

In chapter 8, we study the problem of mining implicit linkage is known as hypotheses generation and its potential to accelerate scientific progress is widely recognized. Almost all of prior studies to tackle this problem ignore the temporal dynamics of concepts. This is limiting because it is known that the semantic meaning of a concept evolves over time. To overcome this issue, in this study, we define this problem as mining time-aware Top-k conceptual bridges and in doing so provide a systematic approach to formalize the problem. Specifically, the proposed model first extracts relevant entities from the corpus, represents them in time-specific latent spaces, and then further reasons upon it to generate novel and experimentally testable hypotheses ($A \rightarrow B \rightarrow C$). The key challenge in this approach is to learn a mapping function that encodes the temporal characteristics of concepts and aligns the across-time latent spaces. To solve this, we propose an effective algorithm that learns precise mapping sensitive to both global and local semantics of input query. Both qualitative and quantitative evaluation are performed on the largest available biomedical corpus. The results obtained substantiate the importance of leveraging the evolutionary semantics of medical concepts and suggest that the generated hypotheses are novel and worthy of clinical trials.

In chapter 9, we present a novel hypothesis generation framework that unearths the latent associations between concepts by modeling their co-evolution across complementary sources of information. More specifically, the proposed approach adopts a shared temporal matrix factorization framework that models the co-evolution of concepts across both corpus and KB. Extensive experiments on the largest available biomedical corpus validates the effectiveness of the proposed approach.

Chapter 10 concludes the dissertation with a discussion of future research directions.

Part I

Specialized Representation Learning for Biomedical Domain

Chapter 2

Learning Knowledge-Powered Word Embedding

2.1 Introduction

Improving distributed representation of words has been at the nucleus of research in Natural Language Processing (NLP) community for a long time [24, 25, 26]. The initial works to tackle this problem were based on the Bag of words (BOW) representation [27, 28], where each individual word is represented as a one-hot vector (i.e., one component in the vector has a value one and rest are zero). However, this representation fails to capture the rich structure of synonyms and antonyms among words. As opposed to this, works like [25, 16, 29] have focused on representing words as continuous low dimensional dense vectors. Vectors of this kind, commonly known as word embeddings, have been shown to capture the implicit semantics of the corresponding words based on the idea of distributional hypotheses (words appearing in similar context have similar meaning) [30].

But does implicit semantics always correlate to meaningful interpretation in real-world? While works like [16, 17] have shown the capability of word embeddings to capture synonyms, the inability to capture meanings under insufficient local context has been pointed out by many researchers [31, 32, 33]. The problem aggravates further in the case of medical domain since one needs to factor in the semantics of the medical terms. Consider the examples in Table 2.1, where we illustrate 2 scenarios. The pair *Heart* and *Blood Vessels* refer to concepts which can be deemed similar at semantics level

Table 2.1: A sample example of medical concept similarity.

Concept 1	Concept 2	Human Correlation	Only Semantic Knowledge	Word Vectors (No Semantic Knowledge)	Word Vectors (With Semantic Knowledge)
heart	blood vessels	0.80	0.65	0.31	0.75
migraine disorders	vascular headaches	0.78	0.30	0.75	0.72

but as shown in the example, the word embeddings are unable to manifest this relation due to lack of local context. However, such scenarios are well captured through external knowledge-bases (KBs) that are curated and maintained by Subject Matter Experts (SMEs). At the same time, relying solely on such curated external coding dictionaries restricts the capability of models to discover terms that have high contextual evidence but are distant in the KBs. Consider the second example in Table 2.1 wherein the pair *Migraine* and *Vascular Headache* are adjudicated to be of low correlation by the external KB despite having high co-occurrence and close semantic proximity in the corpus. Thus, it is necessary to develop a balanced approach such that the model is sensitive not only to the implied semantics but also regulated by the external curated KB. This is the main problem that we are going to address in this paper and in doing so, we provide a systematic approach on how to generate the required prerequisites, cast the objective as a joint optimization and the necessary update rules to solve them for the biomedical domain.

Towards this direction, we leverage MEDLINE[1] , the popular and perhaps the most comprehensive citation repository in the biomedical domain. While this source provides us with the co-occurrence information of the Medical Subject Headings (MeSH terms)[2] required to create its word embeddings, we also utilize MeSH tree codes which serves as the external curated knowledge-base. We provide more details on this in Section 9.3 but it suffices to say at this point that the objective of this paper is to combine these two sources of information to create word embeddings of superior quality as compared to the methods operating on these sources in an isolated fashion. While there have been some works in this area, and even fewer in case of biomedical domain, most of these methods can be categorized to either retrofitting domain knowledge on top of

[1] https://www.nlm.nih.gov/pubs/factsheets/medline.html
[2] https://www.nlm.nih.gov/pubs/factsheets/mesh.html

the learnt word embeddings [34, 35] or one that adopts a similar approach that we describe in this paper. However, it is important to mention that the external coding dictionary that these methods use are usually denoting equivalence/similarity between a pair of concepts and do not capture/measure the similarity between concepts connected indirectly. For instance, the WordNet[36] which is used by authors of [37, 33] provides cognitive synonyms for a word and thus can be used to refine the word vectors to reflect the expert knowledge. However, it cannot quantify the association between any two arbitrarily chosen terms and this is precisely the lacuna in the methodologies using such resources. In our experimental results, we show that such solutions are restrictive in nature and do not suit as well as scale to the biomedical domain and problem setting. Consequently, we leverage the taxonomic structure of the MeSH tree code to compute the distance between any two MeSH terms which is then used as a "regularizer" to refine the word embeddings. It is worthwhile to point out that although we have used MeSH tree and MeSH terms as the inputs, the solution by itself is generic and can be used as a plug-and-play for different tree based external knowledge-bases. Furthermore, compared to existing approaches, the word embeddings we obtain are trained on a relatively small vocabulary. We show that our proposed approach achieves a gain of 13% in terms of Spearman coefficient when compared to state of the art baselines.

To summarize, in this paper, we make the following particular contributions:

1. We propose a new word embedding model for biomedical domain that is not only sensitive to external domain knowledge but intrinsically can handle indirect relationships manifested in it.

2. The vocabulary of the model is small and yet provides a rich representation of the semantic relationships between the medical concepts.

3. Compared to existing best Spearman coefficient result of 0.69, our method achieves a boost of 13% to yield a very high concordance with respect to biomedical experts as well as physicians.

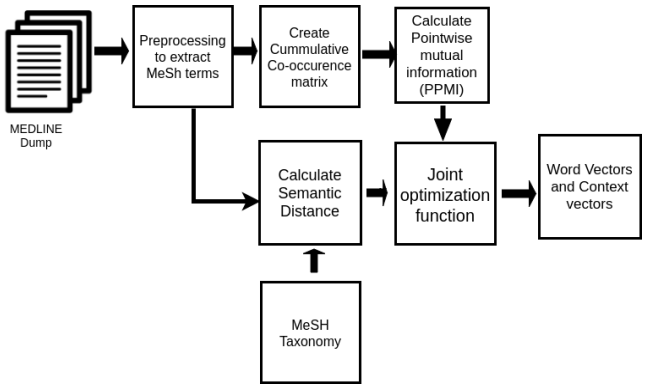


Figure 2.1: Basic Architecture of Proposed Model

2.2 Methodology

2.2.1 Background and Problem definition

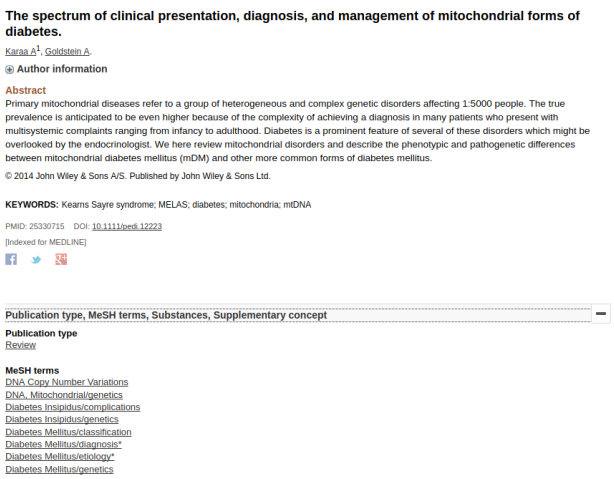
To reiterate, our goal in this paper is to develop external knowledge powered word embedding model for the biomedical domain. The following subsections briefly introduces the corpus and the chosen vocabulary.

MEDLINE

Our corpora is MEDLINE, the largest bibliographic database in biomedicine. At present, it contains more than 24 million references to journal articles from life science and biomedicine. Each article in the corpus contains several attributes such as a unique identifier (PubMed ID), title, abstract, Mesh terms, Publication date, etc. In this work, we use MeSH terms as the source of information and also as our overall vocabulary. As of 2016, there are in total 27,882 MeSH terms and consequently is the size of our vocabulary. The following paragraph introduces the MeSH vocabulary, which is followed by a brief introduction of MeSH tree that serves as our external KB.

Medical Subject headings (MeSH)

Medical Subject Headings (MeSH) are National Library of Medicine (NLM) controlled vocabulary that human experts use to index journal articles in the life sciences domains. Mesh terms are classified into three categories a) Descriptors, b) Qualifiers and c) Supplementary concept records. *Descriptors* represent the conceptual meaning of the



The spectrum of clinical presentation, diagnosis, and management of mitochondrial forms of diabetes.

Karaa A[1], Goldstein A.

Author information

Abstract

Primary mitochondrial diseases refer to a group of heterogeneous and complex genetic disorders affecting 1:5000 people. The true prevalence is anticipated to be even higher because of the complexity of achieving a diagnosis in many patients who present with multisystemic complaints ranging from infancy to adulthood. Diabetes is a prominent feature of several of these disorders which might be overlooked by the endocrinologist. We here review mitochondrial disorders and describe the phenotypic and pathogenetic differences between mitochondrial diabetes mellitus (mDM) and other more common forms of diabetes mellitus.

KEYWORDS: Kearns Sayre syndrome; MELAS; diabetes; mitochondria; mtDNA

PMID: 25330715 DOI: 10.1111/pedi.12223

[Indexed for MEDLINE]

Publication type, MeSH terms, Substances, Supplementary concept

Publication type
Review

MeSH terms
DNA Copy Number Variations
DNA, Mitochondrial/genetics
Diabetes Insipidus/complications
Diabetes Insipidus/genetics
Diabetes Mellitus/classification
Diabetes Mellitus/diagnosis*
Diabetes Mellitus/etiology*
Diabetes Mellitus/genetics

Figure 2.2: Sample document annotated with MeSH terms

article. In this work, we use Descriptors as the unit of representation for documents. Figure 2.2 illustrates a snapshot of citation within the MEDLINE corpus.

External Knowledge-base

In addition to MeSH terms, we also use external expert curated knowledge-base that is organized in a tree hierarchy. Every MeSH term has a corresponding tree code which represents its level of specificity in the tree. As an example, the tree code for concept *Migraine Disorders* is: C10.228.140.546.399.750. Also, in this regard it is important to note that there can be multiple tree codes associated with a single Mesh term. For instance, the MeSH term *Cell Count* in Figure 3.3 has MeSH tree codes E01.370.225.500 and G04.140.

2.2.2 Approach

Traditionally, word embeddings are generated using neural networks with majority of them modelling the objective function as a one trying to predict either the word under consideration based on a context described through a window or the vice-versa [16, 18]. Such methodologies are usually non-interpretable as the neural network acts as a black box. However, recently in [38], the authors proved that the objective function the neural network attempts to solve in case of skip-gram model with negative sampling [16] (word2vec) is the same as matrix factorization of the Shifted Positive Point-wise Mutual

Information (SPPMI) matrix obtained from the co-occurrence matrix of the corpus. In other words, if one creates a co-occurrence matrix of words occurring in a corpus and then calculates the corresponding SPPMI to create a new matrix, the matrix decomposition of the resultant matrix will yield the word and its corresponding context vectors. This proof has subsequently led to many researchers adopting this route over neural network training to obtain word-embeddings; since for reasonably sized matrix, factorization methods like SVD are deterministic in nature. We also adopt a similar approach as it enables us to model the external knowledge-base as one of the components of the objective function. Furthermore, MeSH terms do not have word order information available (a necessity for Neural network based approach) that reinforces the need to adopt the SPPMI based route.

Formally, let us denote D as our text corpus and $V = \{w_1, ..., w_v\}$ as our vocabulary of size $|V|$, where each w_i corresponds to an individual term. Next, we construct a term-by-term $(V \times V)$ co-occurrence frequency matrix, where the rows and columns represent words present in the vocabulary and element represents the raw frequency between them. Now, our goal is to find a dense, low dimensional representation vector $u_w \in R^d$, $d \ll |V|$ for each word $w \in V$. We denote u_{w_i} as the embedding for word w, and d as the embedding dimension. In compact form, U represents the embedding matrix of size $V \times d$, whose i-th row corresponds to the embedding vector of i-th word u_{w_i}.

In the following step, we compute the $|V| \times |V|$ Shifted Positive Point-wise Mutual Information (SPPMI) matrix specific to a corpus D, whose <w,c>-th entry is:

$$SPPMI(w, c) = max(PMI(w, c) - logk, 0) \tag{2.1}$$

where *logk* refers to a global constant. This acts as a prior on the probability of observing a positive example (an actual occurrence of *(w,c)* in the corpus) versus a negative example. A higher value of k indicates that negative examples are more likely. The PMI in the above equation is defined as:

$$PMI(w, c) = \log \frac{\#(w, c).|D|}{\#(w).\#(c)} \tag{2.2}$$

where # (w,c) counts the number of times that words w and c co-occur within a document d over the entire corpus D, #(w) and #(c) denote the total number of times w and c occur in the entire corpus alone. $|D|$ is the total number of word tokens in the corpus.

To represent the conclusion of [38] mathematically, if M represents the SPPMI matrix then,

$$M \approx M_d = U_d \cdot \Sigma_d \cdot W_d^T \tag{2.3}$$

where $M_d = U_d \cdot \Sigma_d \cdot V_d^T$ is the matrix of rank d that best approximates the original matrix M. In such a scenario, the word-vector W_d and C_d is obtained by:

$$\begin{aligned} W_d &= U_d \cdot \sqrt{\Sigma_d} \\ C_d &= W_d \cdot \sqrt{\Sigma_d} \end{aligned} \tag{2.4}$$

where each row in W_d, C_d corresponds to a d-dimensional word-vector and context-vector respectively for a corresponding row (word) in M. While the above discussion illustrates matrix factorization through SVD, the results of the derivation is applicable to any matrix factorization approach when M is viewed as the product of W and C [38]. In our experiments, the value of global constant ($logk$) and dimensionality of the generated MeSH vectors is empirically set to 5 and $d = 200$ respectively. Figure 2.1 pictorially describes the discussion so far.

2.2.3 Proposed Model (MeSH2Vec)

Motivated by the above observation, we formulate our problem as one comprising of three components. The first component is based on the observation where we use the co-occurrence matrix to generate the word embeddings. However, the resulting word embeddings not only needs to satisfy the minimization error in the matrix factorization part but also should have short "distance" in the external knowledgebase (in our case it is the MeSH tree code). This along with the regularizer to prevent over-fitting are the remaining two components of the joint optimization problem. Formally we define it as:

$$\begin{aligned} J = \min \quad & \overbrace{\tfrac{1}{2}||M - WC^T||_2^2}^{\text{Matrix Factorization Component}} \quad + \overbrace{\tfrac{\beta}{2}(||W||_2^2 + ||C||_2^2)}^{\text{Regularizers}} \\ & + \underbrace{\tfrac{\gamma}{2}||M_{dist} - N||_2^2}_{\text{External KB Component}} \end{aligned} \tag{2.5}$$

In expanded form, this can be written as:

$$J = \min \sum_{i,j}^{|V|,|V|} \left(\frac{1}{2}(m_{ij} - \sum_{k=1}^{d} w_{ik} c_{kj})^2 + \frac{\beta}{2} \sum_{k=1}^{d} (w_{ik}^2 + c_{kj}^2) + \frac{\gamma}{2}(n_{ij} - \text{Dist}(\vec{w_i}, \vec{w_j}))^2 \right) \tag{2.6}$$

where m_{ij} refers to the element of matrix M and w_{ik} and c_{kj} refer to the i-th and j-th column of word and context vector respectively. The second part of the objective function (w_k^2 and c_k^2) are the regularizer term to avoid overfitting and β controls the magnitude of word and context vectors. In this work, we set the value of $\beta = 0.01$. As mentioned before, the third part incorporates the prior knowledge into the model, which is regulated by the value of γ. The basic idea being, the word embedding of two words (w_i, w_j) should be closer to each other if they have a smaller semantic distance (higher similarity) in their structured representation. We set the value of $\gamma = 1$ essentially saying giving equal importance to the matrix factorization as well as external KB component. M_{dist} is the distance matrix for word vectors, where the distance adopts the basic Euclidean distance between their word vectors

$$\text{Dist}(\vec{w_i}, \vec{w_j}) = ||\vec{w_i} - \vec{w_j}||_2^2 \tag{2.7}$$

and n_{ij} denotes the element in the i^{th} row and j^{th} column of N (semantic distance matrix), indicating the semantic distance between concepts i and j. The following section provides more details on using this external KB.

2.2.4 Calculating Distance in the External Curated KB

In this section, we describe how to calculate the semantic distance between terms using MeSH taxonomy. As mentioned above, MeSH terms are categorized in a hierarchical fashion. The hierarchical nature of terms can be considered as a "IS A" tree and its structure gives us a concept measure of semantic similarity distance between MeSH terms. Towards this end, there are two aspects of MeSH taxonomy that needs to be considered while calculating the semantic distance: a) The deeply nested structure of MeSH taxonomy (the lower the concepts in the hierarchy the greater shared information they account for) b) One concept in the tree may belong to several sub-categories. To address these issues, in this work, we calculate the semantic distance between MeSH

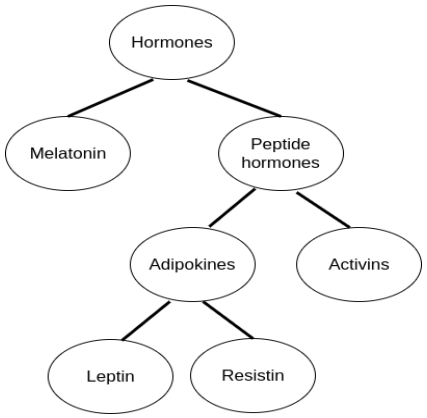


Figure 2.3: Example for calculating semantic distance between concepts in the external tree based KB

terms based on the depth of the least common "subsumer" (LCS) (immediate common parent) and shortest path length between MeSH terms, much akin to [39, 40]. The depth of least common subsumer quantifies the amount of information shared among the concepts in the hierarchy. For the second issue (concepts belonging to multiple categories), we take the minimum semantic similarity distance between the two concepts. The formula for calculating the semantic distance between two concepts is described below:

$$\begin{aligned} n_{ij} = \log_2([\text{distance}(C_i, C_j) + 1] \\ * [D - \text{depth}(lcs(C_i, C_j))]), \end{aligned} \tag{2.8}$$

where distance(C_i, C_j) is the shortest distance between concept C_i, C_j, depth$(lcs(C_i, C_j))$ is the depth of $lcs(C_i, C_j)$, D is the maximum depth of the taxonomy, and $lcs(C_i, C_j)$ is the lowest common subsumer of C_i and C_j.

As an example, in Figure 2.3, the $lcs(leptin, resistin)$ is *adipokines* and its depth is 3, assuming the depth of the root (hormones) is 1. It should be noted that the greater the semantic distance the lower the semantic similarity and vice-versa.

2.2.5 Parameters Update Rules

We take the gradient of our objective function (Equation 2.9) with respect to each of the model parameters w_{i_k}, c_{j_k} and then adopt stochastic gradient descent to update them

(See **??** for derivation). Thus, on each co-occurrence record, this gives us the following closed-form updates:

$$\begin{aligned} w'_{ik} &= w_{ik} + \alpha(-e_{ij}c_{kj} + \beta w_{ik} + \\ &\qquad 2\gamma(||\vec{w}_i - \vec{w}_j||_2^2 - n_{ij})(w_{ik} - w_{jk})), \\ c'_{kj} &= c_{kj} + \alpha(-e_{ij}w_{ik} + \beta c_{kj}), \end{aligned} \tag{2.9}$$

where α is the learning rate. The value of α is empirically set to 0.01.

2.3 Experiments

Having explained the methodological details, we now empirically evaluate, analyze and discuss the proposed model's performance against a variety of biomedical concept similarity/relatedness datasets.

In particular, we attempt to answer the following questions:

1. Are MeSH terms a better source of information for word embedding models in biomedical domain?
2. Does the embedding model based only on co-occurrence/PPMI statistics produce results on par with existing works?
3. Does the augmentation of semantic knowledge to corpus based embedding model improve the overall performance?

2.3.1 Evaluation Datasets

To evaluate the output embeddings on biomedical concept similarity/relatedness task, we borrow evaluation set from [41]. Table 2.2 enumerates the benchmark datasets along with the number of concept pairs that were manually rated by human experts to denote semantic similarity.

- MeSH-1 : The first dataset (MeSH-1) [42] was created by experts from Mayo Clinic and consists of a set of word pairs that are related to general medical disorders. The similarity of each concept pair was assessed by 3 physicians and 9 medical coders. Each pair was annotated on a 4 point scale: practically synonymous, related, marginally, and unrelated. The average correlation between physicians is

0.68 and between experts is 0.78. In our experiments, similar to [39, 34], we found 25 out of 30 concepts pairs as MeSH terms using the latest MeSH dictionary[3] . Also, some of the term pairs in this set were found in the entry terms set of MeSH terms. Every MeSH term has a few corresponding entry terms that are considered to be quasi-synonyms (they are not always exactly synonyms).

- MeSH-2: The second biomedical benchmark (MeSH-2) was introduced in [43]. It consists of a set of 36 word pairs extracted from the MeSH repository. The similarity between word pairs was assessed by 8 medical experts and assigned a score between 0 (non-similar) to 1 (synonyms).

- UMNSRS-SIM: The third dataset (UMNSRS-SIM) was developed by [44] and consists of 725 clinical term pairs whose semantic similarity were determined independently by four medical residents from the University of Minnesota Medical School. Each concept pair was given a score in the range of 0-1600, with higher score implying similar or more related judgments of manual annotators. In our experiments, we mapped these Unified Medical Language System (UMLS) medical concepts to their corresponding MeSH terms and found 218 pairs in the UMNSRS-SIM.

- UMNSRS-REL: Similar to the previous dataset, the fourth dataset (UMNSRS-REL) was also developed by medical residents from the University of Minnesota Medical School [44]. However, the concepts in this dataset were rated for their semantic relatedness rather than similarity. The semantic relatedness score spans the four relatedness categories: completely unrelated, somewhat unrelated, somewhat related, closely related. It should be noted that semantic similarity can be viewed as a special case of semantic relatedness. In our experiments, we mapped these medical concepts to their corresponding MeSH terms and found 221 pairs in the UMNSRS-REL dataset.

2.3.2 Evaluation metric

The way we are going to demonstrate the superiority of our word embeddings is by showing that for the datasets enumerated above, we obtain a better correlation coefficient when compared to the ground truths provided by the experts. Consequently, we

[3] https://www.nlm.nih.gov/mesh/filelist.html

Table 2.2: Summary of datasets used for evaluating semantic similarity/relatedness task for biomedical concept pairs.

Datasets	Concept Pairs
MeSH-1[42]	30
MeSH-2[43]	36
UMNSRS-SIM[44]	218
UMNSRS-Rel[44]	221

use Spearman coefficient as the evaluation metric.

Spearman coefficient (ρ): This metric is used to correlate word pair rankings produced by the proposed method to the ones assigned by expert judgments. The formula for calculating ρ is given in Equation 2.10, where d_i is the difference between the ranks of x_i and y_i, x_i refers to the i^{th} element in the list of human judgments, y_i to the corresponding i^{th} element in the list of semantic similarity computed values, and n is the total number of word pairs. In this work, we use Spearman coefficient to judge the quality of our result.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{2.10}$$

2.3.3 Evaluation Scheme

As the main objective of this section is demonstrate the efficacy of the proposed model, we evaluate it under two conditions - (a) Without incorporating external KB and (b) With incorporating external KB. Through this type of ablation testing, we want to study and quantify not only the benefits of the various components in our system but also understand the scenarios in which a particular component tends to be more useful. As such analyses tend to be dataset specific which can then be generalized, we have added a brief discussion at the end of the results for each dataset.

2.3.4 Results and Discussion

MeSH-1

Table 4.3 presents the Spearman (ρ) coefficient values obtained after applying the proposed model on the first dataset (MeSH-1). This dataset is one of the two dataset (the

Table 2.3: Absolute values of correlation of the five measures relative to human judgments - MeSH-1

Measure	Physician	Expert
Path length[45]	0.627	0.852
Leacock and Chodorow[46]	0.672	0.856
Wu and plamer[47]	0.652	0.794
Choi and Kim[48]	0.560	0.724
Nguyen[39]	0.672	**0.862**
Yu et al.[34]	0.696	0.665
MeSH2Vec (without prior knowledge)	**0.817**	0.76
MeSH2Vec (prior knowledge)	**0.836**	0.801

Table 2.4: Absolute values of correlation of the five measures relative to human judgments- MeSH-2

Measure	Human expert
Baseline	0.55
Aouicha et al[41]	0.724
MeSH2Vec (without prior knowledge)	**0.79841**
MeSH2Vec (prior knowledge)	**0.81542**

Table 2.5: Correlation values relative to human judgments for UMNSRS-Similarity

Measure	Human expert
Pyysalo et al.[49]	0.549
Chiu et al.[50]	0.652 (N/A)
Munneb et al.[51]	0.52 (N=462)
Pakhomov et al.[52]	0.62 (N=449)
Mcinnes et al.[53]	0.66 (N=401)
MeSH2Vec (without prior knowledge)	**0.74** (N=218)
MeSH2Vec (prior knowledge)	**0.75** (N=218)

other being MeSH-2) where the original concepts were MeSH terms and we did not have to perform any mapping. Consequently, results in this dataset assumes more importance than the latter as it give a fair one-to-one comparison with the baselines. As it

Table 2.6: Correlation values relative to human judgments for UMNSRS-Relatedness

Measure	Human expert
Pyysalo et al.[49]	0.506 (N/A)
Chiu et al.[50]	0.601 (N/A)
Munneb et al.[51]	0.45 (N=465)
Pakhomov et al.[52]	0.58 (N=458)
Mcinnes et al.[53]	0.49 (N=401)
Aouicha e al.[41]	0.634 (N/A)
MeSH2Vec (without prior knowledge)	**0.70** (N=221)
MeSH2Vec (prior knowledge)	**0.73** (N=221)

can be observed from the table, the proposed model (MeSH2Vec) outperforms others and achieves the highest correlation with Physician's judgments. In terms of Expert judgments, even though we do not make it the top, the difference between our approach and the baselines is comparable (considering they perform good only in one of the two columns).

Discussion: For comparing our results with previous works, we extracted results reported by [39] and [34]. The first five techniques [45, 46, 47, 48, 39] in Table 4.3 are the ontology only techniques (they purely rely on taxonomic information), while the fifth is a retrofitting technique based on both context vectors and semantic lexicons. Since they rely on the taxonomical details, which are usually prepared by experts, they have high concordance in that evaluation criterion. However, from the perspective of users, i.e., physicians, much of these approaches perform poorly. To analyze the reason for the difference between the perception of coders and physicians, we would like to recapitulate the discussion in [54]: The medical coders were more sensitive to the hierarchical classification thereby being more inclined to the concept of (taxonomic) similarity whereas physicians seemed to represent a more general concept of (taxonomic and non-taxonomic) relatedness. The rationale seems plausible as the ontology based measures (such as [39]), presumed better at capturing taxonomic similarity have better correlation with Coders whereas the context vector based methods (such as the proposed method) that is sensitive to both taxonomic and contextual information have better correlation with physicians. The proposed model outperforms others and achieves the highest correlation with physician's judgments and yet maintains a comparable result

with expert's judgment. Since the experts are usually medical coders following some form of guidelines and not physicians, we need a mechanism that balances the need of both. We believe our methodology does that.

Another observation is that the more recent technique of retrofitting (it fits semantic lexicons in a post-processing step) introduced by Yu et al. [34] obtains improvement over ontology-only techniques based on Physician correlation. However, they have low correlation with the coders. Perhaps, the reason lies in the inability of this method to fully leverage the hierarchical structure of MeSH tree. In contrast, the proposed model obtains significant improvement over the aforementioned work in both physician and coder's judgment. Analyzing results further, in our perspective, the boost in performance is because the proposed model apart from capturing the implicit similarity between words via co-occurrence/PPMI statistics also integrates evidence of multiple taxonomic paths between concepts and relative densities of their taxonomical branches present in the MeSH hierarchy.

MeSH-2

Table 4.4 shows the correlation values obtained for the Spearman (ρ) coefficient for MeSH-2 dataset. Unlike the first dataset (MeSH-1), there are no results reported by ontology only techniques on this dataset. In order to have a baseline, a correlation score was calculated by only considering the taxonomic information of MeSH hierarchy and the formula introduced in Equation 2.8. The proposed model obtains the highest correlation as compared to baseline and existing works.

Discussion: More recently, Aouica et al. [41] evaluated their intrinsic information content based similarity measure on this dataset and reported their best correlation score as 0.724. The proposed model improves upon this baseline. Upon further analyses of results, we observe that the limited performance of this method is due to its nature to ignore the contextual information present in plain text.

UMNSRS-SIM and UMNSRS-REL

Tables 2.5 and Table 2.6 show the correlation scores of our proposed model on University of Minnesota Semantic Relatedness (UMNSRS) datasets. In these tables, N refers to the number of concept pairs that we were able to successfully map between MeSH and UMLS concepts (recall that these datasets are based on UMLS concepts and not all

UMLS concept need to have a corresponding MeSH term). Such form of analysis was also done by other researchers in past and for comparison, we report the correlation scores of these existing works along with the number of pairs they were able to map to this dataset. Note that they do not use MeSH terms and consequently the numbers are not the same.

Discussion: The substantial improvement of our approach over existing baselines should be attributed to the joint exploitation of local contextual evidence and relational information from taxonomy. As an example, concept pairs such as "*appendicitis*" and "*peritonitis*" which despite being semantically related (they both are related to Intra-abdominal infections) have diverse context in the corpus. Thus, the methods based solely on co-occurrence/PPMI statistics have limited performance. However, the addition of taxonomic evidence from MeSH hierarchy refines their embeddings to be closer to each other in euclidean space. Apart from that, another added advantage of the model is the relatively lower size of vocabulary. In comparison to previous works such as [55], which require a larger token size (93,095,323), our proposed model attains higher or on par performance on a smaller dictionary (27,882). We believe the reason for this is the high quality input, thanks to MeSH terms, that provides an accurate representation of concepts.

2.4 Related Works

In the past few years, a series of works [25, 56, 16, 57] have applied deep learning techniques to learn distributed word representation. These methods have shown dramatic improvement in the performance of several NLP tasks. For instance, Collobert et al. [24] proposed a neural network that learns a unified word representation suited for tasks such as parts of speech tagging, named-entity recognition and semantic role labeling. Similarly, Socher et al. in [58] improved the performance of sentiment analysis task and semantic relation classification using recursive neural network. More recently, Mikolov et al. [16, 18] proposed two efficient neural network models: a) Continuous bag-of-words model (CBOW) and b) word2vec for learning word representation. These models are unsupervised in nature and trained on large text corpora. Particularly, these models maximize the log likelihood of each word given its context words within a sliding window. They have been shown to capture analogical relations and improved performance in various evaluations [59, 23, 60, 61].

In the biomedical domain, recent years have seen some early attempts towards applying word embedding model for bioNLP tasks. Munnet et al. [51] trained both the Skip–gram and CBOW models over the PubMed Central Open Access (PMC) corpus with approximately 400 million tokens. On the task of semantic similarity and relatedness, they report that Skip-gram model (word2vec) performed the best for the task of semantic similarity, on the other hand, none of the models outperformed others in the semantic relatedness task. Chiu et al. [50] performed analysis on the effect of input corpora, architecture and hyperparameter setting (negative sample size, sub-sampling, minimum-count, learning rate, vector dimension, and context window size) on the quality of embeddings. In their results, they report the values of some of the influential hyperparamater, 10 (negative sample size), 1e-4 (sub-sampling), 0.05 (learning rate), 200 (vector dimension), and 30 (context window size) for word2vec and conclude that the size of corpora does not affect the quality of word embeddings. Similar to the aforementioned work, a more recent study [62] examined the effect of recency, size and section of biomedical publication (abstract/full-text) data on the performance of word2vec. They reported that the models trained on recent datasets did not boost the performance and as compared to the full text articles bodies, abstracts excel in accuracy.

Despite the prominent role played by above works in highlighting the salient aspects of biomedical embedding models, they did not make any model level innovation and at their core followed the distributional hypotheses [30]. In other words, the embeddings generated by these model suffer for those words that are infrequent or unseen during training, such as domain-specific words. To circumvent this problem, it is necessary to incorporate domain knowledge. It is known that the biomedical domain has abundant amount of taxonomic and relational information stored in form of vocabularies and ontologies, however little attention has been paid to integrate them into the embedding model itself. In this work, the proposed model collectively learns the word representation by exploiting both the co-occurrence statistics from plain text and semantic evidence from the domain knowledge. In this regard, it is worthwhile to point out that there have been some works in NLP domain to include the prior knowledge [55, 63, 64, 32]. As an example, [37] proposed a simple but effective method to encode relational knowledge. In particular, they proposed a new learning objective that incorporates both a neural language objective and a semantic prior knowledge objective. Similarly, [33] proposed an alternate method to encode semantic knowledge via ordinal constraints. However, these models have mostly been limited to general domain text and their architecture

does not fit the structural representation of taxonomies present in biomedical domain.

Perhaps, the work that is very closely related to ours is a retrofitting method proposed by [34] that incorporates semantic lexicons into the vector representation as a post-processing step. However, as raised by the authors themselves this model does not completely leverage the hierarchical structure of MeSH vocabulary. In contrast to all of these aforementioned works, in this paper, we present a general method that jointly exploits both contextual information and coherency of taxonomical knowledge to learn better word representation for biomedical application.

Chapter 3

Dynamic Word Embedding

3.1 Introduction

Understanding the semantics and intent behind a text is a core task in the field of Natural Language Processing (NLP) [26, 25]. As a precursor to any application with real-world significance, this task has garnered much attention from many researchers leading to the development of various models with distinct assumption on the structure and organization of text [56, 30, 65]. Lately, practitioners in the community have become interested in applying deep learning inspired language models - word embedding models [16, 18] - to learn the distributed representation of words. Apart from being scalable, these models in conjunction with the distributional hypothesis (popularly explained as a word is known by the company it keeps) have been shown to capture the implicit semantics in a much better fashion. Notable examples of word embedding models that have accomplished significant improvement in the performance for several NLP tasks include word2vec [16] and Glove [17]. Despite considerable advances achieved, a major drawback of these models lie in their assumption of a static world. Simply put, these models assume that the semantic meaning of a concept remains the same over the period of time. This is problematic because it is known that the domains in general are dynamic with concepts periodically acquiring new semantic sense and losing old ones [13]. As a simple illustration, consider the word *Intelligence* - during the early 1960's, its meaning used to be associated with concepts such as "war", "opponent" and "experts"; however, lately (2018) it is more often associated with concepts like "artificial intelligence" and "cognitive reasoning". Such drifts in the meaning of a word is observed

in almost all the domains, however, its effect is especially prevalent in domains such as biomedicine, where some new facts emerge and some are rendered obsolete every now and then [13]. Capturing these semantic drift over time is crucial to understanding the dynamics of medical concepts and the evolution of overall medical knowledge. Motivated with this, in this study, we consider the problem of detecting semantic shifts in meaning and usage of medical concepts over a given time frame based on the text data. More specifically, the objective is to characterize the semantic change incurred across time frames and encapsulate them into the learned representation of medical concepts.

As mentioned before, studies [17, 16] in the past have investigated the problem of learning distributed representation of words; however, models that are sensitive to the dynamic nature of domain are scarce. Though a few recent studies [66, 67, 68] have attempted to tackle this problem, more or less, they follow a two-step process: a) compute static word embeddings in separate time-frames separately, and b) then find a way to align these embeddings across time-frames. Adopting such two-step procedure has a few inherent issues. First, as embedding models are known to have stochastic nature of initialization, training them separately might result in produced embeddings being less interpretable [69]. Second, as theses approaches only consider two time frames (instead of all) of embeddings at each alignment phase, they jeopardize the quality of embeddings learned [69, 70]. To mitigate these aforementioned issues, in this study, we systematically formulate this problem of learning distributed representation of medical concepts from the sequential text. In doing so, we discuss the procedure to generate the required prerequisites, cast the objective as a joint optimization and provide the necessary update rules to solve them. The proposed model, Dynamic Word Embedding for Medicine (DWE-Med), basically captures the temporal dynamics by relying upon the principles of statistical co-occurrence and temporal smoothing. Essentially, the core idea of proposed technique is to compute the word embeddings and alignments jointly, through solving one overall optimization problem. In addition, the alignment strategy over all time slices (instead of two) enables the model to learn embeddings of higher quality. Lastly, the temporal smoothing step encourages the embeddings to vary smoothly over time, thereby, easing the interpretation and visualization of results. As will be shown in the experiments later (See Section 9.4), the dynamic embeddings learned using this approach facilitates an intuitive trajectory visualization of concepts by tracking neighboring words across times. For example, a concept *homosexuality* has a trajectory of *substance-related disorder* $\rightarrow$ *gender identity* (See Figure 3.1).

In addition to presenting a dynamic language model, we also demonstrate the effectiveness of dynamic embeddings by studying their role in two novel use cases from the biomedical domain. These use cases are: a) hypotheses generation [23, 7, 22] and b) ontology expansion [71]. The goal of first use-case (i.e., hypotheses generation) is to find implicit linkage between previously disjoint topics of interest. In other words, given two concepts (A, C) that have no known direct connection, the objective is to find B terms that connects them in a novel way. These new connections are hitherto unknown and therefore called hypothesis ($A \rightarrow B \rightarrow C$). Our intent behind this use-case is to elucidate the importance of dynamic embeddings in the process of evolutionary association formation process, wherein, the concepts with a potential of forming a connection iteratively come closer to each other in each time-stamp. Apart from the potential of aiding scientists in the process of formulating novel hypothesis, this task also benefits other related areas of biomedical research such as drug-drug interaction and biomedical QA [72]. While the first use-case provides insights into the capability of DWE-Med in standalone applications, the choice of second use-case illuminates its importance for downstream applications. Towards this end, we choose another related problem from biomedical domain, namely, ontology expansion. The objective of ontology expansion is to predict the branches on an ontology that will undergo expansion in near future. The motivation behind this task is to automate the process of ontology evolution thus easing its current practice of manual maintenance by the subject matter experts. Furthermore, as our goal in this study is to explore the utility of word embeddings in a dynamic setting, the task of ontology evolution provides a relevant use-case.

Note that this paper is an extended version of our previous study [23]. In this article, we adopt and extend the methodological innovation of [23] to propose a general framework for dynamic word embedding and further study its applicability for a completely novel use-case. In particular, our contributions can be summarized as:

1. In this study, we propose a general framework for dynamic word embedding, that is capable of modeling the gradual semantic evolution of medical concepts over time. As a supplement, these embeddings allow us to track and visualize the evolutionary trajectories of various concepts, thereby, achieving fine interpretability.

2. The capability of dynamic embeddings to encode both implicit semantics and evolutionary behaviors of medical concepts enables us to reveal previously unknown associations in the medical domain.

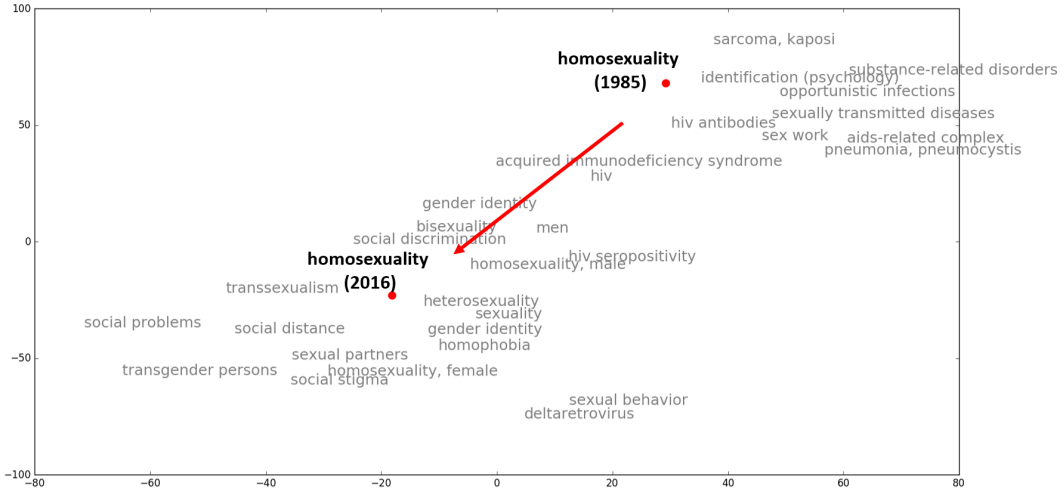


Figure 3.1: Two dimensional projection of word embeddings for the concept *homosexuality* and its trajectory visualization using t-Distributed Stochastic Neighbor Embedding (t-SNE)

3. A novel RNN based deep architecture to predict the expansion of ontology concepts is proposed. The superiority of predicting ontology expansions from an evolutionary perspective is validated in the experiments.

4. Extensive experimentation is conducted on bio-medical corpus spanning more than hundred years. New detailed analysis and discussion are presented for both the chosen use-cases.

3.2 Methodology

This section first describes the proposed DWE-Med model in detail, and then elucidates their applicability in two biomedical use cases. As our focus in this study is to learn the temporally sensitive dense representation of words, we first need a text corpus collected across time. This corpus is then split into distant time scopes, and a co-occurrence matrix of medical concepts for each time period is created. The co-occurrence statistics thus obtained is used to learn evolutionary characteristics of medical concepts. More specifically, first all the documents are aggregated to the granularity of five years, e.g., 1900-1904, 1905-1909, 1910-1914 and so on. Then, for each time slot t, a co-occurrence matrix $X^{(t)}$ of medical concepts is constructed to capture the co-occurrence patterns,

wherein, each entry $X_{ij}^{(t)}$ denotes the number of times that the i^{th} concept co-occurs with the j^{th} concept in the same article.

3.2.1 Evolutionary Word Embeddings

Given T time-stamped concept co-occurrence matrices $\{\boldsymbol{X}^{(1)}, ..., \boldsymbol{X}^{(t)}, ..., \boldsymbol{X}^{(T)}\}$, the semantics and evolutionary patterns of each medical concept are carried implicitly within those matrices. In the following sub-sections, we describe how our model learns these semantics and evolutionary patterns of medical concepts in an optimal manner. Section 3.2.1 first describes the method to learn static word embeddings and then Section 3.2.1 generalizes this to the dynamic setting.

Static Word Embeddings

To learn the static embeddings, we adopt and extend a popular word embedding model, namely, GloVe [17]. The main motivation behind adoption of GloVe is its capability to leverage the benefits of both global matrix factorization (e.g., Latent Semantic Analysis) and local context window methods (e.g., Skip-gram) simultaneously. Concretely, the model achieves this by training explicitly on the non-zero elements in a word-word co-occurrence matrix, instead of training on the entire matrix that are generally sparse. Motivated with this unique aspect of GloVe to take the best of both worlds, we choose to follow this line of research. We assume that the co-occurrence information described the semantics of a concept, i.e., its context information. As an example, consider two concepts $i = male$ and $j = female$, their relationship can be examined by studying the ratio of their co-occurrence probabilities with other probe terms, k: for terms k like *brain* or *carbon*, that are related to both male and female, or to neither, we expect the co-occurrence probability ratio $P(k|i)/P(k|j)$ to be close to one; for terms k more related to female than to male, say $k = pregnancy$, the probability ratio $P(k|i)/P(k|j)$ should be small; in contrast, for terms more related to male than to female, the ratio should be large. This assumption suggests that the probability ratio $P(k|i)/P(k|j)$ depends on two target terms i, j and one probe term k. By adopting the vector difference and the dot product of the embeddings, the linear structures of the embedding space can be captured and modeled via:

$$F((\boldsymbol{w}_i^{(t)} - \boldsymbol{w}_j^{(t)})^\intercal \widetilde{\boldsymbol{w}}_k^{(t)}) = \frac{P(k|i)}{P(k|j)}, \tag{3.1}$$

where $\boldsymbol{w}^{(t)} \in \mathbb{R}^d$ are embeddings at time stamp t and $\widetilde{\boldsymbol{w}}^{(t)} \in \mathbb{R}^d$ are the context embeddings at time stamp t, respectively. $\boldsymbol{w}_i^{(t)}$ is used when term i works as a target term, and $\widetilde{\boldsymbol{w}}_i^{(t)}$ is used when term i works as a probe term. Given the term-term co-occurrence matrix at time t, $X^{(t)}$, $P(k|i)$ is empirically set as $P(k|i) = X_{ik}^{(t)}/X_i^{(t)}$, where $X_i^{(t)} = \sum_m X_{im}^{(t)}$ is the number of times any concept co-occurred with another concept i at time t. Thus, by taking F as the exponential function and adding biases, a simplification over Equation 3.1 is obtained:

$$\boldsymbol{w}_i^{(t)\intercal}\widetilde{\boldsymbol{w}}_k^{(t)} + b_i^{(t)} + \widetilde{b}_k^{(t)} = \log(X_{ik}^{(t)}), \tag{3.2}$$

where $b_i^{(t)}$ and $\widetilde{b}_k^{(t)}$ are biases associated with term i and k at time t. Considering that the term co-occurrence matrix $X^{(t)}$ is very sparse, static embeddings for time stamp t can be learned via a weighted least squares regression:

$$J^{(t)} = \sum_{i,j=1}^{V} f(X_{ij}^{(t)})(\boldsymbol{w}_i^{(t)\intercal}\widetilde{\boldsymbol{w}}_j^{(t)} + b_i^{(t)} + \widetilde{b}_j^{(t)} - \log(X_{ij}^{(t)}))^2, \tag{3.3}$$

where f is a weighting function for each entry in the co-occurrence matrix. As suggested in GloVe [17], f is set as:

$$f(x) = \begin{cases} (x/x_{max})^\alpha & \text{if } x < x_{max} \\ 1 & \text{otherwise} \end{cases}. \tag{3.4}$$

The reason for choosing this weighting function is the following: a) For large values of x, $f(x)$ is relatively small. This property prevents the frequent co-occurrences from being overweighted. b) As it observed, $f(x)$ is a non-decreasing function and thus the rare term are not over-weighted.

DWE-Med

The previous subsection introduced the procedure to learn static word embeddings from an independent term co-occurrence matrix. Now, given a time sequence of term co-occurrence matrix $\{\boldsymbol{X}^{(1)}, ..., \boldsymbol{X}^{(t)}, ..., \boldsymbol{X}^{(T)}\}$, we aim to learn dynamic word embeddings that evolves smoothly over time.

Figure 3.2 shows the framework of our DWE-Med model. The learned dynamic word embeddings at time t must account for both their semantics which are carried by the current term co-occurrence matrix and their historical evolutionary trajectories. At

each time stamp t, we add a distance constraint to each medical concept which prevents the embedding from drifting too far from its historical location:

$$O^{(t)} = \sum_{i,j=1}^{V} f(X_{ij}^{(t)}) \Big((\boldsymbol{w}_i^{(t)\intercal} \widetilde{\boldsymbol{w}}_j^{(t)} + b_i^{(t)} + \widetilde{b}_j^{(t)} - \log(X_{ij}^{(t)}))^2 + \beta I^{(t)}(i) l(\boldsymbol{w}_i^{(t)}, \boldsymbol{w}_i^{(t-1)}) \Big),$$

where β is the parameter controlling the damping to the historical embeddings, $I^{(t)}(i)$ is a indicator function, and $l(\boldsymbol{w}_i^{(t)}, \boldsymbol{w}_i^{(t-1)})$ measures the distance between term i's current location in the embedding space $\boldsymbol{w}_i^{(t)}$ and its historical location $\boldsymbol{w}_i^{(t-1)}$. $I^{(t)}(i)$ indicates if term i has occurred in history:

$$I^{(t)}(i) = \begin{cases} 1 & \text{if term } i \text{ has occurred before time } t \\ 0 & \text{otherwise} \end{cases}. \tag{3.5}$$

A large number of distance measurements can be used as $l(\boldsymbol{w}_i^{(t)}, \boldsymbol{w}_i^{(t-1)})$, such as cosine distance, but since we aim to learn smooth evolutionary trajectories of medical concepts, we adopt the Euclidean distance between the current embeddings and historical embeddings:

$$l(\boldsymbol{w}_i^{(t)}, \boldsymbol{w}_i^{(t-1)}) = ||\boldsymbol{w}_i^{(t)} - \boldsymbol{w}_i^{(t-1)}||^2. \tag{3.6}$$

In practice, β is set to a small value, so the damping to the history is very weak. At time stamp $t = 1$, we define $I^{(0)}(i) = 0$. We put the embedding shift constraint $l(\boldsymbol{w}_i^{(t)}, \boldsymbol{w}_i^{(t-1)})$ only on word embeddings, because context embeddings might need to change its scale frequently as the scale of the co-occurrence matrices changes. Thus, the overall objective function of our DWE-Med model is as follows:

$$O = \sum_{t=1}^{T} O^{(t)} = \sum_{t=1}^{T} \sum_{i,j=1}^{V} f(X_{ij}^{(t)}) \Big((\boldsymbol{w}_i^{(t)\intercal} \widetilde{\boldsymbol{w}}_j^{(t)} + b_i^{(t)} + \widetilde{b}_j^{(t)} - \log(X_{ij}^{(t)}))^2 + \beta I^{(t)}(i) l(\boldsymbol{w}_i^{(t)}, \boldsymbol{w}_i^{(t-1)}) \Big). \tag{3.7}$$

Equation 3.7 enforces that the DWE-Med model learns dynamic embeddings which vary smoothly over time. On each occurrence of a concept, its corresponding embedding

is regulated not to drift too far from its historical location. Thus, the higher term frequency, the larger regulation to be stable over time. This is consistent with the law of conformity of language evolution – 'rates of semantic change scale with a negative power of word frequency' [66]. DWE-Med efficiently shares information across the time domain, which allows us to feed the time-stamped data sequentially in steps. The dynamic embeddings thus produced are able to capture both the implicit semantics of concepts and also track their temporal change.

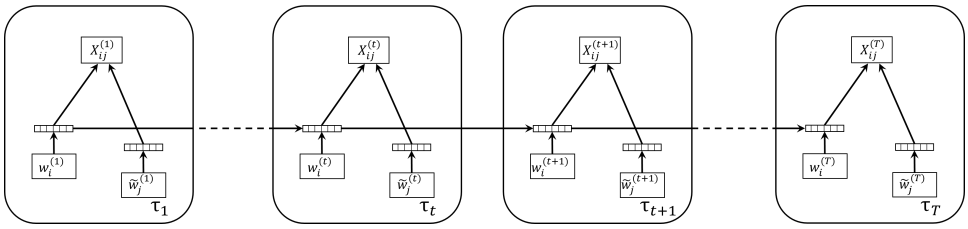


Figure 3.2: Framework of DWE-Med. T time slices of data are connected via dynamic word embeddings.

3.2.2 Parameter Inference

We take the gradient of DWE-Med objective (Equation 3.7) with respect to each of the model parameters $\{\boldsymbol{w}_i^{(t)}, \widetilde{\boldsymbol{w}}_j^{(t)}, b_i^{(t)}, \widetilde{b}_j^{(t)}\}$ and then adopt stochastic gradient descent to update them. Thus, on each co-occurrence record, this gives us the following closed-form updates:

$$\begin{aligned}
\boldsymbol{w}_i^{(t)} &\leftarrow \boldsymbol{w}_i^{(t)} - \eta * 2f(X_{ij}^{(t)}) \Big((\boldsymbol{w}_i^{(t)\intercal} \widetilde{\boldsymbol{w}}_j^{(t)} + b_i^{(t)} + \widetilde{b}_j^{(t)} \\
&\qquad - \log(X_{ij}^{(t)})) \widetilde{\boldsymbol{w}}_j^{(t)} + \beta I^{(t)}(i)(\boldsymbol{w}_i^{(t)} - \boldsymbol{w}_i^{(t-1)}) \Big), \\
\widetilde{\boldsymbol{w}}_j^{(t)} &\leftarrow \widetilde{\boldsymbol{w}}_j^{(t)} - \eta * 2f(X_{ij}^{(t)}) \Big((\boldsymbol{w}_i^{(t)\intercal} \widetilde{\boldsymbol{w}}_j^{(t)} + b_i^{(t)} + \widetilde{b}_j^{(t)} \\
&\qquad - \log(X_{ij}^{(t)})) \boldsymbol{w}_i^{(t)} \Big),
\end{aligned}$$

$$\begin{aligned}
b_i^{(t)} &\leftarrow b_i^{(t)} - \eta * 2f(X_{ij}^{(t)})(\boldsymbol{w}_i^{(t)\intercal} \widetilde{\boldsymbol{w}}_j^{(t)} + b_i^{(t)} + \widetilde{b}_j^{(t)} - \log(X_{ij}^{(t)})), \\
\widetilde{b}_j^{(t)} &\leftarrow \widetilde{b}_j^{(t)} - \eta * 2f(X_{ij}^{(t)})(\boldsymbol{w}_i^{(t)\intercal} \widetilde{\boldsymbol{w}}_j^{(t)} + b_i^{(t)} + \widetilde{b}_j^{(t)} - \log(X_{ij}^{(t)})),
\end{aligned}$$

where η is the learning rate. Note that the vectors are randomly initialized at the beginning, and then are subsequently updated in the later time units. Further details on the parameters are provided in Section 9.4.

Having explained the core idea behind dynamic embedding, we now illuminate the applicability of DWE-Med in two biomedical use-cases (See Section 3.2.3 and Section 3.2.4).

3.2.3 Use-Case I (Hypotheses generation)

Given two previously disjoint terms i and j along with a cut-off time-stamp t, the task is to identify high confident bridge terms k that will connect them in the future (after t). In doing so, all the concepts present in the vocabulary (besides i and j) are considered as our candidate bridge concepts. To find promising bridge concepts among the possible candidates, we filter and rank these intermediary k terms using the following criteria: (1) term k's cosine similarity with i and j at current time stamp t: k should be close to both input terms to be a bridging term; (2) the evolutionary trajectories: k is favored if there is a growing association trend between k and i, j; (3) the generality of term k: we prefer informative terms to generic terms. Therefore, the intermediary terms k are ranked according to:

$$s(k|i,j,t) = \text{sim}(\boldsymbol{w}_k^{(t)}, \boldsymbol{w}_i^{(t)}, \boldsymbol{w}_j^{(t)})\Delta(\boldsymbol{w}_k, t)\text{trd}(\boldsymbol{w}_k, \boldsymbol{w}_i, \boldsymbol{w}_j, t), \tag{3.8}$$

where $\text{sim}(\boldsymbol{w}_k^{(t)}, \boldsymbol{w}_i^{(t)}, \boldsymbol{w}_j^{(t)})$ denotes k's cosine similarity with i and j at time t. To penalize terms that are close to only one input term but far away from the other input term, we adopt F1 cosine similarity score as $\text{sim}(\boldsymbol{w}_k^{(t)}, \boldsymbol{w}_i^{(t)}, \boldsymbol{w}_j^{(t)})$:

$$2\frac{\text{cos-sim}(\boldsymbol{w}_k^{(t)}, \boldsymbol{w}_i^{(t)}) * \text{cos-sim}(\boldsymbol{w}_k^{(t)}, \boldsymbol{w}_j^{(t)})}{\text{cos-sim}(\boldsymbol{w}_k^{(t)}, \boldsymbol{w}_i^{(t)}) + \text{cos-sim}(\boldsymbol{w}_k^{(t)}, \boldsymbol{w}_j^{(t)})}.$$

$\Delta(\boldsymbol{w}_k, t)$ reflects the generality of k till time t as defined in Equation 3.9 (See Section 3.2.3), and $\text{trd}(\boldsymbol{w}_k, \boldsymbol{w}_i, \boldsymbol{w}_j, t)$ is the association trend between k and i, j up until time t, defined as:

$$\exp\left(\text{acs}(\boldsymbol{w}_k^{(t)}, \boldsymbol{w}_i^{(t)}, \boldsymbol{w}_j^{(t)}) - \text{acs}(\boldsymbol{w}_k^{(a)}, \boldsymbol{w}_i^{(a)}, \boldsymbol{w}_j^{(a)})\right),$$

where $\text{acs}(\boldsymbol{w}_k^{(t)}, \boldsymbol{w}_i^{(t)}, \boldsymbol{w}_j^{(t)})$ stands for the average cosine similarity between them, and a denotes the first time stamp they appeared.

Demoting Generic Concepts

One challenge in the problem of hypotheses generation is to discern between informative and generic terms. Simply put, generic terms (also known as stopwords in NLP domain),

such as humans and animals tend to frequently co-occur with a wide variety of other terms, and thus tend to have high association score with most of the terms. However, these generic terms are to be demoted when ranking bridge concepts, as they are not informative. The conventional approach to tackle this issue is to utilize certain heuristic rules such as removing the concepts which appeared more than 10,000 times in the entire corpus [73]. However, such heuristics lack clear rationale behind them. To do this in a more effective manner, we use the metric proposed in Equation 3.9 to penalize highly general terms.

$$\Delta(\boldsymbol{w}_i, T) = \frac{1}{N_i} \sum_{t=1}^{T-1} I^{(t)}(i) * \text{cos-dist}(\boldsymbol{w}_i^{(t)}, \boldsymbol{w}_i^{(t+1)}), \tag{3.9}$$

where N_i is the number of time slices a term i occurred. The basic idea is to penalize the terms in accordance with their semantic stability. Highly frequent or general words tend to have more stable meaning over time [66], as a result, the average cosine distance change over time for these terms (calculated by Equation 3.9) is lower. This cause the overall candidate score of generic terms to decrease.

3.2.4 Use-Case II (Ontology Expansion)

In this task of Ontology expansion, our focus is to demonstrate the utility of dynamic embeddings as a feature. To briefly recapitulate the problem: Given the past ontology versions and a related corpora, the task is to predict the branches in the ontology that will undergo expansion in the future. Figure 3.3 provides a conceptual description of the problem statement. While the proposed methodology to tackle this problem is applicable to any tree-based Ontology, for the sake of illustration, we choose Medical Subject Headings (MeSH) [1] from the biomedical domain.

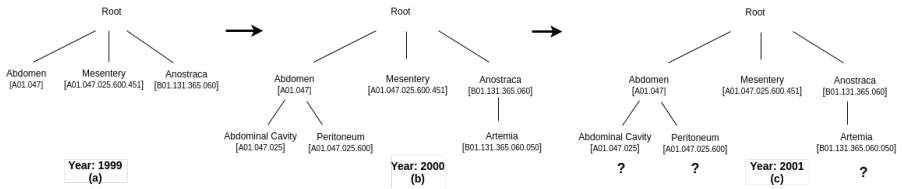


Figure 3.3: Snapshot depicting evolution of Ontology

[1] https://www.nlm.nih.gov/mesh/mbinfo.html

As mentioned in the introduction, our objective is to tackle this problem from an evolutionary perspective. This requires us to define and quantify the semantic change of ontology concepts. In more detail, the semantic change of a concept is defined as the total semantics directly associated with a concept or indirectly through its immediate neighbors. For example, consider the medical concept "choline" that was initially associated only to "amines". However, the continued research over time with "amino alcohols", "quaternary ammonium compounds" and other neighboring terms increased the semantic density of "choline" by including near-synonyms and subsets. We assume that this increasing trend of semantic density gives rise to the need for higher specificity of "choline" in the ontology, i.e., the possible expansion of concept "choline" in the ontology. To encapsulate this semantic change as a feature, we use the dynamic embeddings of ontology concepts (Further details in Section 3.2.4).

As an auxiliary source of information, we also use other two type of features (See Section 3.2.4 and Section 3.2.4).

Temporal Entropies.

While dynamic embeddings provides features representing semantic change, it is also important to incorporate usage diversity of concepts. Simply put, if a term turns more and more polysemous over time, it has a high chance to expand. This polysemous level of a term can be measured by the entropy score from information theory that is based on its contextual diversity in the corpus. As the focus of this work is on capturing temporal dynamics, we adapt the measure to incorporate the time component. The temporal entropy for term i at time t is described as below:

$$e_i^{(t)} = \sum_j p(c_j^{(t)}, w_i^{(t)}) \log \frac{p(w_i^{(t)})}{p(c_j^{(t)}, w_i^{(t)})},$$

where $c_j^{(t)}$ is a context word of $w_i^{(t)}$ at time t, $p(c_j^{(t)}, w_i^{(t)}) = X_{ij}^{(t)} / \sum_k X_{ik}^{(t)}$ denotes the occurrence probability of $w_i^{(t)}$ with its context $c_j^{(t)}$, and $p(w_i^{(t)}) = \sum_j p(c_j^{(t)}, w_i^{(t)})$ denotes the occurrence probability of $w_i^{(t)}$.

Temporal Structural Features.

This feature captures the temporal characteristics of concepts from their taxonomic structure (a portion of MeSH hierarchy is shown in Figure 3.3). We use six such

structural features for each time stamp t. Specifically, we use three static features, namely, *minDepth*, *maxDepth*, *siblings* and three dynamic features, namely, *tempMinDepth*, *tempMaxDepth* and *tempSiblings*. *MinDepth* and *MaxDepth* refer to the minimum and maximum depth of a medical concepts at time t. *Siblings* refer to the number of medical concepts that share at least one common parent with term i. For the remaining three dynamic features, they measure the difference between time stamp t and previous time stamp $t-1$:

$$f^{(t)}_{i,tempSiblings} = \frac{f^{(t)}_{i,siblings} - f^{(t-1)}_{i,siblings}}{f^{(t)}_{i,siblings}}.$$

Since all our features are temporal in nature, the problem blends itself into a sequence modelling task. Towards this end, we propose a Recurrent Neural Networks (RNN) based deep architecture namely, Evolutionary MeSH Expansion (EME) to model the evolutionary characteristics of medical concepts. A complete pipeline of the proposed framework is illustrated in Figure 3.4.

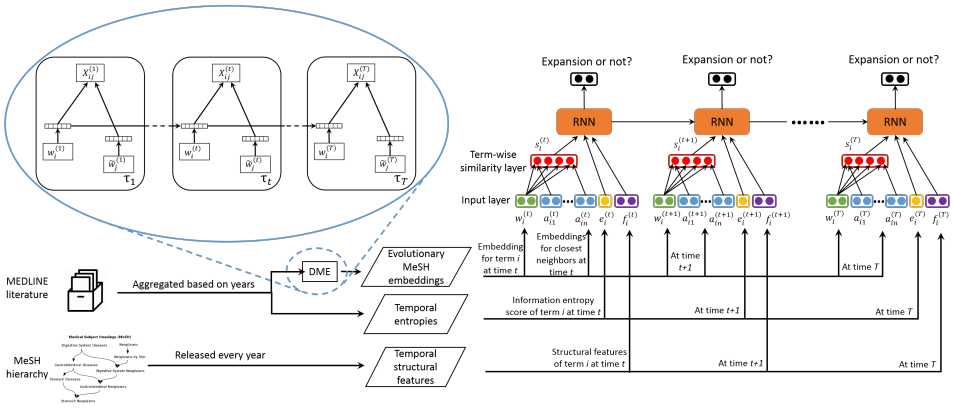


Figure 3.4: Proposed framework for Ontology Expansion.

Evolutionary MeSH Expansion (EME)

The goal of the proposed model is to predict the expansion label (yes or no) for a term at time T. Given a training sample in the form of {term i, time T}, features from time t_0 to T are utilized to predict whether term i will expand at time T or not, as illustrated by the right half of Figure 3.4. In practice, we set $t_0 = T - 10$ to incorporate term i's previous 10 years' evolution until time T.

RNNs provide an elegant way of modeling sequential data. In Figure 3.4, "RNN" could be any RNN variants, such as Long Short Term Memory (LSTM) [74] and Gated Recurrent Unit (GRU) [75]. In our implementation, we use GRU to deal with the vanishing gradient problem. Based on the temporal features obtained, at each time stamp t, the RNN takes dynamic medical concept embeddings $\boldsymbol{W}^{(t)}$, temporal entropies $\boldsymbol{e}^{(t)}$, temporal structural features $\boldsymbol{f}^{(t)}$ as well as the hidden state $\boldsymbol{h}^{(t-1)}$ from the previous time slot as the input. To measure the semantic change of term i, the first layer is a semantic similarity layer, where the k^{th} entry denotes the similarity between term i and its k^{th} closest neighbor in the embedding space:

$$s_{i,k}^{(t)} = \text{cos-sim}(\boldsymbol{w}_i^{(t)}, \boldsymbol{a}_{i,k}^{(t)}),$$

where $\boldsymbol{a}_{i,k}^{(t)}$ denotes the embedding for term i's k^{th} closest neighbor. Cosine similarity is adopted as the similarity measurement due to its simplicity and effectiveness. Other similarity measurements such as Euclidean distance could also be utilized here. We then employ a concatenation layer to combine the semantic similarities with the temporal entropies and structural features. The hidden state at time t is calculated as follows:

$$\boldsymbol{h}^{(t)} = \text{GRU}(\boldsymbol{h}^{(t-1)}, [\boldsymbol{s}^{(t)}; \boldsymbol{e}^{(t)}; \boldsymbol{f}^{(t)}]).$$

Hidden state $\boldsymbol{h}^{(t)}$ is subsequently fed to a softmax layer to produce the expansion prediction $\hat{\boldsymbol{y}}^{(t)} \in \mathbb{R}^2$ for time t:

$$\hat{\boldsymbol{y}}^{(t)} = \text{softmax}(\boldsymbol{U}_o \boldsymbol{h}^{(t)} + \boldsymbol{v}), \tag{3.10}$$

where $\boldsymbol{U}_o$ and $\boldsymbol{v}$ are the weight matrix and biases to be learned. Based on Equation 3.10, our objective function can be calculated as the cross-entropy between ground truth expansion label $\boldsymbol{y}^{(t)}$ and prediction expansion label $\hat{\boldsymbol{y}}^{(t)}$:

$$L = -\frac{1}{T - t_0} \sum_{t=t_0}^{T} (\boldsymbol{y}^{(t)\intercal} \log(\hat{\boldsymbol{y}}^{(t)}) + (1 - \boldsymbol{y}^{(t)\intercal}) \log(1 - \hat{\boldsymbol{y}}^{(t)})).$$

3.3 Experiments

The focus of this section is to demonstrate the efficacy of dynamic embeddings for both the chosen use cases. In our experiments, we use MEDLINE[2] as our main corpora. The

[2] https://www.nlm.nih.gov/pubs/factsheets/medline.html

latest dump (2017) contains more than 24 million articles, primarily, from the domain of life sciences and biomedicine. Every article contains a unique identifier (PMID), title, abstract, publication date and Medical Subject Headings (MeSH) terms. As a unit of representation for articles, we choose MeSH terms. MeSH terms are the special keywords assigned by subject matter experts to each article in MEDLINE. Since these terms are selected by subject matter experts based on the full text of articles, it is safe to assume that they represent the conceptual meaning of an article. Furthermore, the choice of MeSH terms (over other sources of representation such as title/abstract) has certain specific benefits: a) previous studies [9, 6] have shown that the use of concepts from plain title/abstract introduces noise into the system and is also computationally expensive, b) several articles pre-1990 have limited or no abstract content and c) being assigned by human experts they are concise. These particular benefits make MeSH terms a good choice for unit of representation. On average around 13 MeSH terms are assigned to each article in MEDLINE [76]. To obtain the dynamic word embeddings[3] , as suggested by a few previous studies [17, 77], we empirically set $\alpha = 0.75$, $\beta = 0.01$, $x_{max} = 100$, $\eta = 0.05$ and run DME for 100 iterations on each time slice. The value of dimension is also empirically set to $d = 200$. Based on conclusions provided in existing word embedding literature [16, 50], the dimensionality of embeddings is determined by examining its performance on word similarity and relatedness tasks [50].

3.3.1 Use-Case I (Hypotheses generation)

To assess the effectiveness of our model, we perform both qualitative and quantitative evaluation. The qualitative evaluation determines the extent to which our approach is capable of rediscovering the known knowledge, while the quantitative evaluation is intended to analyze the overall quality of results.

Fish-oil (FO) and Raynaud's Disease (RD): In 1986, Swanson [5] investigated the research question of "role of dietary fish oils in treating patients with Raynaud's syndrome". By manually inspecting literature belonging to Fish oils and Raynaud's disease respectively, he found that Raynaud's disease is worsened by high *blood viscosity*, high *platelet aggregation*, *Vasoconstriction*, and the ingestion of Fish oils reduced these phenomena.

In our results, it can be seen that both *platelet aggregation* and *blood viscosity* are

[3] The source code of dynamic embeddings is made available at `https://www.dropbox.com/sh/pi28cwxg46f5xy9/AAAtII95M3ypJw5aQtL1-Q0ja?dl=0`.

found at rank 8 and 11 respectively. With regards to this, it is worthwhile to note that many rediscovery approaches consider it a success if they find *platelet aggregation* in their list of intermediates [10]. Furthermore, other important connections besides the ground truth include *'fattyacids, essential'* and *'vasodilation'*.

Magnesium (MG) and Migraine Disorder(MIG): Swanson [78] proposed eleven bridging connections between Magnesium and Migraine Disorder. These important connections are: *epilepsy, serotonin, prostaglandins, platelet aggregation, calcium antagonist, type A personality, vascular tone and reactivity, calcium channel blockers, spreading cortical depression, inflammation, brain hypoxia and substance P*. Unlike the previous test case, we are unable to achieve high recall. Nonetheless, we obtained important connections such as *epilepsy*, *calcium channel blockers*, *adenosine triphosphate*, etc. With regards to this, it should be noted that previous research indicates this to be a difficult test case [6].

Somatomedin C (SMC) and Arginine (ARG): Somatomedin C (SMC) (also known as Insulin-like Growth Factor I (IFG1)) is a growth regulating peptide and Arginine is an important amino acid. They both were found to be associated to each other through the means of growth hormones such as *somatotropin* and *somatostatin*. Growth hormones tend to influence SMC and ARG in turn triggers the secretion of growth hormones.

As it can be observed, in our results, somatotropin is ranked number 1 and somatostatin is found in top K. In comparison to the prior works [6] which use ad-hoc semantic types to get these results, our model finds them in a completely automated way.

Indomethacin (INN) and Alzheimer Disease (AD): An important research question of whether Alzheimer Disease (AD) - a progressive disease that destroys memory and other important functions, can be treated with an inflammatory agent - Indomethacin (INN) was explored during 1990's. Researchers reported that connections such as Acetylcholine, Membrane fluidity to be important bridging pathways. In our findings, similar to previous test case, Acetylcholine is ranked 1. Although Membrane fluidity was not ranked in top K, its derivatives were found at higher ranks. An interesting observation worth mentioning is regarding the term *nitric oxide* (Rank=3). Although not yet experimentally proven, several papers identified nitric oxide as important for understanding alzheimers [6]. Moreover, during 2000-2001, there were works [79] that depicted strong influence of nitric oxide in both Alzheimer's disease and Indomethacin.

Schizophrenia (SZ) and Calcium - Independent Phospholipase A2 (CI-PA2): Schizophrenia is a disease that affects the person's ability to think, feel and concentrate. It has been found to be elevated in patients suffering from SZ. After synthesizing independent works of [80] and [81], Swanson and Smalheiser hypothesized oxidative stress to be the key connecting term. In our findings, we were able to obtain oxidative stress indirectly through receptors, adrenergic (PMID: 3820966). Also, much alike to the previous test case, our top ranked term (glutamates) is found to be heavily investigated for its influences in treating Schizophrenia (PMID: 20686195) during more recent years.

Overall, the proposed model was able to identify a majority of true connections at top ranks. Next, we illuminate on how the availability of dynamic embedding facilitates visualization of intuitive trajectories.

Evolutionary Trajectories

As our medical knowledge develops, the semantics (medical properties) of medical evolve, for example, the finding of a new medication or a new cause to a specific disease would probably result in their medical properties getting more similar. This semantic evolution is reflected as evolutionary trajectories of medical concepts in the embedding space. Consider the classic example of *Fish Oils (FO) – Blood Viscosity (BV) – Raynaud's Disease (RD)*, Figure 3.5 shows the two-dimensional projection of the dynamic embeddings and their evolutionary trajectories using t-Distributed Stochastic Neighbor Embedding (t-SNE) [82]. As it can be noticed, initially in 1953, all three concepts were at different positions, but, as the research over these topics increased in parallel, their implicit semantics started getting closer, making them very close to each other in 1983. This evolutionary behavior eventually in 1986 led to their co-occurrence for the first time in a research article.

Quantitative evaluation

In the previous subsection, we discussed the capability of proposed model to predict novel associations much ahead of their real discovery time. The focus of this section is to measure the overall quality of ranked set. To do so, we need a ground truth. Unfortunately, there is no standard ground truth available, therefore, a "supposed" ground truth (based on the documents published after the cut-off date) is generated.

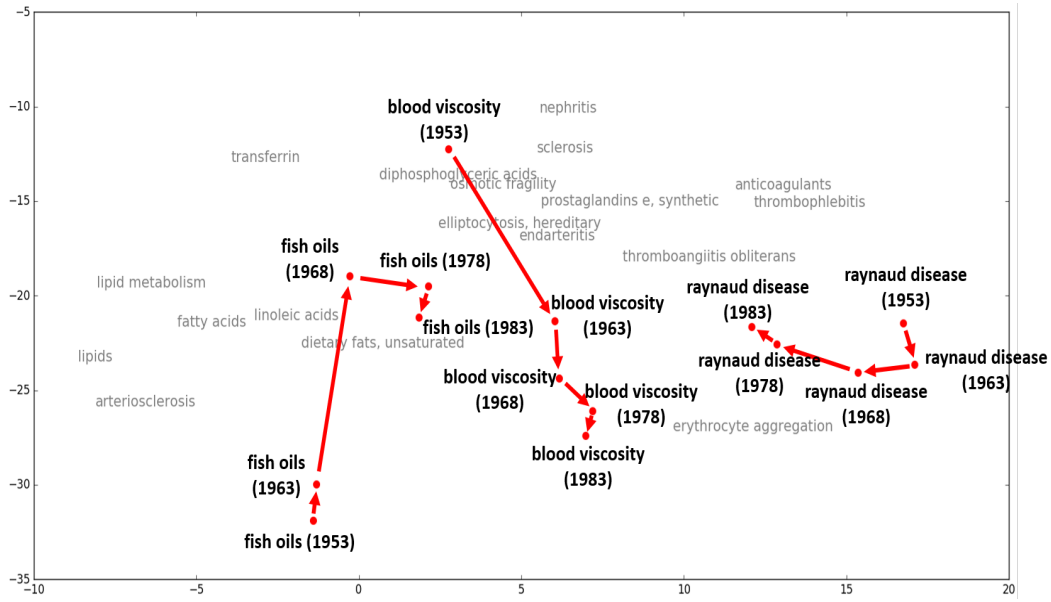


Figure 3.5: An example of the evolutionary behavior of MeSH embeddings.

As an example, consider test case "FO-RD" whose cut-off year is 1985, so based on the documents in 1985-2016, the ground truth intermediate terms k are ranked according to:

$$gt(k) = \frac{\#(k, \text{“}FO\text{”}) + \#(k, \text{“}RD\text{”})}{\#(k)}, \tag{3.11}$$

where $\#(i, j)$ is the number of times terms i and j co-occured, and $\#(i) = \sum_j \#(i, j)$. Hence, the ranked hypotheses can be evaluated by measuring the Spearman's rank correlation with the ground truth set. We compare the proposed DME model with three baselines:

- **Graph [83]**: Graph is a distributional-graph theoretic approach and we implemented our own version of it. This approach uses a combination of graph-based global and local measures to rank the bridging terms between a given pair of input concepts.

- **Static**: Static refers to the standard word2vec embeddings [16], trained on the entire corpus without respecting the temporal information.

- **Transformed-CBOW (Trans)** [68]: In Trans, the embeddings are first trained separately by factorizing PPMI matrix for each year t, and then transformed by

optimizing a linear transformation matrix which minimizes the distance between $w(t)$ and $w(t+1)$ for the semantically stable words (i.e., those words whose semantic meaning remains relatively stable over the period of time).

The comparison results on the five test cases are reported in Table 3.1, 3.2, 3.3, 3.4 and 3.5. The first column of each table is calculated on the entire ground truth ranked set. As it can be observed, the proposed model consistently outperforms the baselines. Analyzing the results further, we gain several insights. First, the improved result obtained by "Static" over "Graph" demonstrates the importance of modelling the problem in latent space. We speculate the reason for this improvement lies in the ability of proposed model to capture implicit connections in an effective manner. Second, leveraging the temporal component of biomedical domain proves crucial. This is validated by the improvement in results for both dynamic embedding approaches (Trans and DME) as compared to "Static" alone. Notably, the reasonable performance for "Trans" lies in its ability to align latent space using frequent words. As the semantic of meaning of such frequent words (example: "humans", "male", "female") tend to remain relatively stable over time, they act as good "anchors" to bridge distinct latent spaces [68]. While this provides an important insight, yet, as it can be observed the performance of "Trans" is still lower than that of the proposed model. We believe this is because of the high quality alignment of embeddings achieved by the proposed model. In this regard, one important point to note is that the proposed framework performs "smoothing" over all the words, whereas, the "Trans" performs alignment using only a set of frequent words. To example with an illustration, consider the case shown in Figure 3.5, in this example the semantic meaning of medical concept "blood viscosity" evolves from nephritis (1953) $\rightarrow$ erythocyte aggregation (1968) $\rightarrow$ thromboangitiis obliterions (1983) over the period of time. Such uniform nature of evolution cannot be precisely captured by alternate dynamic embedding methods such as "Trans" because they do not perform any kind of "smoothing".

For other prior works, as they were performed under different experimental settings and a complete ranked set is difficult to obtain, a direct comparison with their results cannot be performed.

Table 3.1: Spearman's Correlation for FO-RD.

Methods	Top 1505	Top 500	Top 100	Top 20
Graph	0.236	0.142	0.086	-0.266
Static	0.423	0.429	0.440	0.055
Trans	0.426	0.429	0.450	0.060
DME	**0.430**	**0.430**	**0.460**	**0.066**

Table 3.2: Spearman's Correlation for MIG-MG.

Methods	Top 3976	Top 1500	Top 300	Top 200
Graph	0.203	0.035	-0.023	0.013
Static	0.351	0.186	0.161	0.152
Trans	0.354	0.193	0.169	0.163
DME	**0.357**	**0.201**	**0.178**	**0.174**

3.3.2 Use-Case II (Ontology Expansion)

In this section, we present the experimental results for the second use case of Ontology expansion. In our literature review, we found that there is almost no (besides one exception [71]) existing work on ontology expansion optimized for MeSH. This sole prior work cannot handle the newly added MeSH nodes as those nodes have no occurrence before, thus a direct comparison with their results cannot be performed. In order to compare and evaluate our results, we implemented the following supervised learning models as our baselines: Support Vector Machine (SVM) [84], Random forest [85] and Logistic regression [86]. Before we plunge into the details of quantitative evaluation, we first illustrate how the model facilitates in capturing the semantic density change of medical concepts.

Table 3.3: Spearman's Correlation for INN-AD.

Methods	Top 5351	Top 2500	Top 500	Top 100
Graph	**0.188**	0.036	0.051	0.023
Static	0.163	0.139	0.224	0.230
Trans	0.165	0.141	0.232	0.235
DME	0.168	**0.144**	**0.239**	**0.239**

Table 3.4: Spearman's Correlation for IGF1-ARG.

Methods	Top 7599	Top 4000	Top 400	Top 300
Graph	0.266	0.185	0.063	0.063
Static	0.307	0.192	0.169	0.172
Trans	0.313	0.195	0.182	0.181
DME	**0.319**	**0.197**	**0.196**	**0.183**

Table 3.5: Spearman's Correlation for SZ-CI,PA2.

Methods	Top 519	Top 100	Top 50	Top 20
Graph	0.121	-0.244	-0.034	0.058
Static	0.317	0.362	0.176	0.202
Trans	0.319	0.392	0.197	0.312
DME	**0.327**	**0.412**	**0.247**	**0.373**

Semantic Change of Medical Concepts.

The semantic change of a medical concept refers to the amount of semantics that the concept carries or around that concept, manifested as the compactness in its neighborhood in the embedding space. The evolving body of scientific literature causes the addition of new semantically related terms, and the semantic association of a medical concept to drift with other pertinent concepts, thereby increasing its semantic density. For illustration, consider Figure 3.6, which shows the semantic density change of medical

concept "p38 Mitogen-Activated Protein Kinases". The red dot represents this medical concept and the black dots represent its top 10 closest neighbor terms in the embedding space. As it can be observed, over the passage of time, the region around the concept become more and more dense. We believe this escalation in semantic density gives rise to the need for a higher level of specificity, thus resulting in concept expansion.

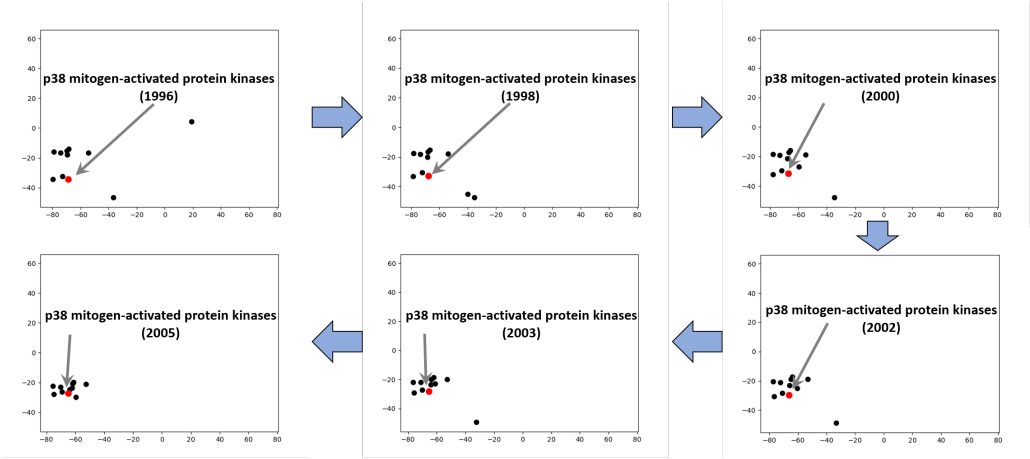


Figure 3.6: Evolving semantic density of a medical concept.

Experimental Setup.

Our dataset for evaluation is the MeSH releases from 2001 to 2016. Firstly, it should be noted that although MeSH exists since 1963, it is only since 1999 that it has been systematically maintained [71]. The training samples are generated year-wise. Now, as the problem is a classification task, for each year, the dataset is first split into positive samples and negative samples. A positive sample {term i, time t} means term i expands at time t, while a negative sample {term i, time t} means term t doesn't expand at time t. The number of negative samples is much larger than the number of positive samples. Therefore, we randomly select an equal number of negative samples as the positive samples to make the dataset balanced. Then we perform 10-fold cross-validation.

Results and Discussion.

In Table 3.6, we report the micro-averaged Precision, Recall and F1-score for the proposed framework and baselines. Note that for this problem, we calculate Precision, Recall and F1 for both the positive and the negative class. Positive class refers to

the case when a term i expands at time t and negative class refers to the case when the term i does not expand at time t. The final reported results are the average of these two classes. Analyzing results further, figure 3.7 shows the yearly performance of the models over the last 15 years in terms of F1-score. The reported results show that EME is able to predict the concept expansion with F1-score reaching 0.90 in the year of 2014. Now, as the main focus of this work is to demonstrate the efficacy of the temporal features used, we examine our results under three conditions - a) with only literature based temporal features, i.e., dynamic MeSH embeddings and temporal entropies (EME-literature), b) with only temporal structural features (EME-structure) and c) with the combination of both (EME). Through this kind of analysis, we intend to study and quantify not only the benefits of the various feature set in our system but also understand the scenarios in which a particular feature set tends to be more useful.

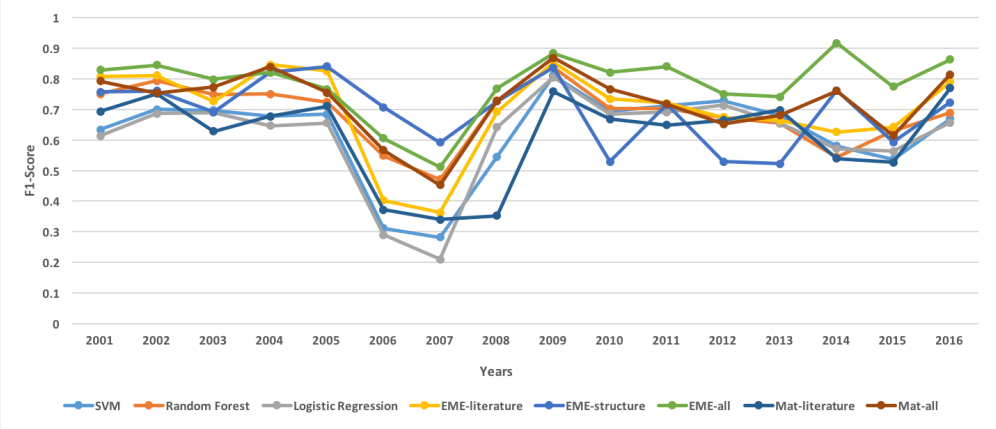


Figure 3.7: F-score comparison of proposed model with baselines

For EME-literature, it can be observed from Table 3.6 that it performs better than SVM and logistic regression. Its performance is comparable to the best baseline Random Forest. For EME-structure, we implement the model using only temporal structural features. As shown in the results, EME-structure surpasses all the baselines. Upon further inspection, we observe that the continuous evolution trend of the topology plays a crucial role to simulate concept expansion. Finally, we probe the results by aggregating both the structural and the literature based features. The result of EME achieves the best performance. Interpreting the reported results further from a broader perspective, we believe that the boost in performance is due to the ability of the proposed model to capture temporal dynamics of medical concepts. The escalation in semantic density

Table 3.6: Prediction results

	Precision	Recall	F1
SVM	0.576	0.576	0.575
Random Forest	0.701	0.678	0.687
Logistic Regression	0.574	0.574	0.573
EME-literature	0.640	0.622	0.631
EME-structure	0.696	0.696	0.694
Trans-ALL	0.727	0.721	0.724
EME	**0.733**	**0.745**	**0.739**

and usage diversity gives rise to the need for a higher level of specificity, thus analyzing all the temporal features together yields a better result. As a consequence, it can be substantiated that the temporal dynamics of a medical concept play a crucial role in their probable expansion.

Having demonstrated the importance of leveraging the temporal component for this particular task, next, we are interested in examining the quality of "dynamic embeddings" itself. To do so, we generated temporal embeddings using an alternate dynamic embedding approach (i.e., transformation based approach [68]) and utilized them as features for this task. The results obtained are reported in Table 3.6. Notably, "Trans-ALL" achieves better performance than the baselines such as SVM, Random Forest and Logistic regression. This observation reinforces the importance of leveraging temporal dynamics for tasks such as Ontology expansion. Nevertheless, as it can be seen from the Table 3.6, the overall highest performance is achieved by the proposed model (EME). Similar to the previous use-case of hypotheses generation, we speculate that this is because of the higher quality of alignment achieved by the proposed model; thereby, encapsulating the temporal dynamics at a much granular level.

3.4 Related Works

Improving distributed representation of words has been an important problem in the research area of NLP [25]. As a consequence of advances made in the research area of deep learning, recently, a series of studies [25, 56, 16] have applied neural network inspired models to learn the distributed representation of words. Collobert et al. [56] proposed a deep learning based framework that aims to learn distributed representation of words

useful for task such as named-entity recognition and semantic role labeling. Similarly, [58] proposed a recursive neural tensor network to improve the performance in the task of sentiment analysis. More recently, studies such as [16] and [17] proposed two scalable language models, namely, word2vec and Glove. These models are unsupervised in nature and trained on large text corpora. Generally speaking, these models maximize the log likelihood of each word given its context words within a sliding window. Apart from capturing implicit semantics at a finer level they also provide special analogical features such as *vec("ibuprofen") - vec("pain")* $\approx$ *vec("treats")*. While these models made substantial strides in this area of study, yet, almost all of them ignored the temporal dynamics of concepts. In this study, we aim to learn the time-aware distributed representation of words by modelling the problem in a dynamic setting.

To mitigate the limitations of a static domain, a few recent studies [68, 67] have taken initial steps towards incorporating the temporal aspect in their language models by adopting a two-step approach: a) compute the static word embeddings in each time-frame independently and b) find a way to align them. A key challenge in these approach is to achieve the alignment. To do so, [67] found a linear transformation of words between any two time slice by solving a least squares problem of k nearest neighbor words. Another study [68] utilized semantically stable words - those words whose meaning do not change between the two time slices - as anchor points to compute the linear transformation. [66] imposed the transformation to be orthogonal, and solve procrustes problem between every two adjacent time slices to perform alignment. While these studies have substantiated the importance of considering time-specific semantics and are able to capture the temporal change in an effective way (particularly in "online settings"), however, they do not perform any kind of smoothing. To address this, in [23], we proposed a joint optimization based approach, wherein, the embeddings and alignments are learned simultaneously. Furthermore, we perform smoothing over the learning process that allows us to leverage the correlation between embeddings at successive time-stamps and learn "smooth" evolutionary trajectories. More recently, few studies such as [70, 69] also proposed a joint modelling based approach, nevertheless, we differ from them in several aspects. First, the probabilistic approach proposed in [70] requires a "sequence" information of words present in the natural language text. However, in the current problem of interest (and perhaps several other real-world scenarios), the word sequence information is not available and the medical concepts are assigned to their respective articles simply as 'bag-of-words'. Thus, to handle this distinct problem

setting, in this study we adopted a bilinear regression based model that exploits the statistical co-occurrence information and generates quality word embeddings. Second, in comparison to [69], our core objectives differ. In this study, we aim to understand the semantic evolution of medical concepts and capture their true meaning in vector representations that are useful for a multitude of knowledge discovery tasks in the medical domain.

In another line of research, few studies from the research area of temporal topic modeling [87] and temporal information retrieval [88, 89] recognized and attempted to tackle this problem of quantifying semantic change. In [87], the authors proposed a probabilistic approach to develop and analyze the temporal evolution of topics. Similarly, [88] proposed a LDA inspired model to capture the low-dimensional structure of change of data over time. More recently, [89] used brownian motion to model the temporal change of latent topics through a time-stamped collection of documents. Our study has a connection to them in a sense them we too are interested in studying the semantic evolution of concepts, nonetheless, our objectives differ. We aim to learn the time-aware vector representation of concepts from large-scale sequential text, whereas, their goal is to detect emerging topics.

Chapter 4

Learning Interpretable Word Embedding

4.1 Introduction

Modelling the lexical semantics behind a word has acquired significant interest in the recent years [25, 18, 90]. As a consequence of advances made in the research area of deep learning, more recently, practitioners in the community have applied neural network inspired language models (commonly known as word embedding models [18]) to model the latent structure present in the text, and produced more nuanced form of word representations. Simply put, these word embeddings models learn to generate dense, continuous, low-dimensional vectors representation of words from raw, unannotated corpora in a completely unsupervised manner. Such succinct form of representation thesedays have become the "de-facto" word representation for a multitude of downstream bioNLP tasks such as disease diagnosis [91], drug re-purposing [92] and hypotheses generation [23, 22].

Despite their considerable success and widespread adoption, a drawback of these word embedding models lie in their inability to provide meaningful interpretation of the individual embedding dimensions. This is problematic because even though we can comprehend the underlying mathematical principles of such models, it is still important to understand what exactly do these dimensions signify? What kind of properties are being (and *not* being) captured by these dimensions? As a simple illustration, consider the example of medical concepts "Insulin" and "Diabetes mellitus" shown in Figure 4.1.

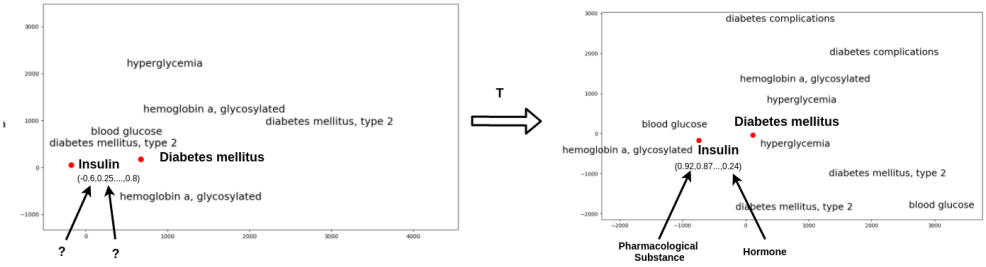


Figure 4.1: The original word embedding space (left) and the transformed embedding space (right).

As it can be observed, the current word embedding models can capture the semantic proximity between these concepts, yet, they cannot answer questions like: "To what extent the medical concept insulin captures the property of being a *pharmacological substance* or a *hormone* ?". In contrast, with the aid of proposed transformation technique (Figure 4.1), we can precisely answer such questions. The main advantage of having such form of post-hoc reasoning is that these interpretable representations might not only aid in generating explainable answers to the sensitive downstream medical tasks such as disease diagonosis [91], but also provide us with keen insights into the nature of state-of-the-art embedding models themselves. Motivated with these speculations, in this study, we consider the problem of improving the interpretablity of words embeddings learned over a particular text corpora.

Unlike numerous studies done on generating vector representations, literature on learning interpretable word embeddings is relatively scarce: [93, 94, 95]. In general, the key idea of these prior studies to improve the interpretability of vector representations is by inducing sparsity in the word vector dimensions [93]. Arguably, these studies made substantial advances, however, they still have a few inherent drawbacks. First, either these models cannot be learned over pre-trained word vectors available from the widely used embedding models such as Word2Vec [18]/GloVe [17] or they produce vectors with much higher dimensions. Second, these studies did not attempt to elucidate the particular conceptual notion (property) being carried within these individual dimensions.

To mitigate these aforementioned issues, in this study, we systematically formulate this problem of improving the intepretability of word embeddings. Basically, the core idea of the proposed model is to leverage upon the rich categorical/taxonomic knowledge present in the biomedical domain and learn a transformation matrix being sensitive to

them. As the available categorical knowledge is manually curated and maintained by subject-matter-experts, our conjecture is that the interpretability of word embeddings in terms of these human-defined categories will reflect more proximity to the human level interpretations. Towards this end, we propose a novel framework that first infers the vector representation of categorical concepts and then learns a transformation matrix that is able to transform the original word embeddings to a new space where these aforementioned categorical concepts act as their basis. Besides, the learning of transformation matrix is performed in such a way that the expressive features of original vectors are retained.

In this study our contributions can be summarized as:

1. We propose a novel framework for interpreting word embeddings, that is capable of transforming any pre-trained word embedding to a new space such that the hidden conceptual meaning of individual dimensions are revealed. To the best of our knowledge, we are among the first to study the interpretability of word embedding in the medical domain.

2. By leveraging upon the principles of dictionary learning and exploiting the categorical knowledge present in the biomedical domain, the proposed technique learns to infer the categorical representations at a granular level.

4.2 Related Work

Improving interpretability of word embeddings has been an active area of study over the past few years [96, 95, 93]. The initial study [96] proposed a non-negative matrix factorization based technique, namely, Non-Negative Sparse Embedding (NNSE) to learn the interpretable embeddings. While this study elucidated the importance of studying interpretability of word embeddings, yet, they were shown to suffer from memory and scale issues. To alleviate this, [97] proposed to learn interpretable embeddings in an online manner. In doing so, their key idea was to adopt a neural network approach to learn the word embeddings, and then employ an adaptive gradient descent to accelerate their convergence.

Building upon the ideas of aforementioned studies, [93] proposed a principled sparse coding technique to improve the interpretability of word vectors. Basically, they utilized

sparse coding in a dictionary learning setting to obtain longer, sparser and overcomplete vectors. A potential drawback of this study is that it produces vectors of very high dimensions. More recently, another study [98] adopted l_1 regularization into their learning objective to induce sparsity and learned interpretable vectors. In general, the central notion behind these sparsity inducing techniques is that they aim to increase the sparseness in vectors, that then leads to better separability, thereby, improving the interpretability. While crucial insights were gained from these aforementioned studies, they still did not focus on explicating the precise conceptual/categorical meaning being carried within the individual dimensions. In this study, by relying upon the principles of categorical theory [99] and correspondingly exploiting the rich categorical knowledge present in the medical domain we attempt to study the interpretability of word embedding dimensions at a more granular level.

The work much akin to ours is a recent study done by [95]. In this study, the authors proposed to rotate the original vector dimensions in such a way that the rotated vectors are interpretable. While close in spirit, we differ from this study in two aspects. Firstly, the objectives are different. We aim to study the interpretability of words embedding in the medical domain by leveraging upon the categorical knowledge. Secondly, our problem is more difficult in a sense that the we aim to particularly illuminate the implicit conceptual notion remaining hidden within these individual dimensions.

4.3 Overview of Proposed Model

Recall that the input to our system is a set of pre-trained word vectors of medical concepts, and the goal is to learn a transformation matrix that projects the input embeddings to a new space wherein the transformed embeddings are both interpretable and retain their original expressive features.

To accomplish our first objective (interpretability), we focus on exploring the principles of category theory [99] and aim to interpret the embeddings in terms of these categories. Such categories in the biomedical domain refer to a broad subject themes that provide a consistent categorization of the medical concepts [100]. These categories in addition to possessing a conceptual meaning also have dictionary definitions associated with them. By taking advantage of this expert knowledge, we infer their categorical representations. These inferred categorical representations then further act as the basis

for our new space. Once this new space is defined, we then learn a transformation matrix from the original embedding space to this new target space. This transformation matrix in particular allows us to achieve interpretability for the input embeddings in the transformed space.

Next, to achieve our second objective (i.e., retaining the expressive features present in the pre-trained vectors), a form of orthogonal constraint is imposed on the learned transformation. Such form of imposition allows us to minimize the possible loss of information; thereby, aiding us to achieve our second objective of retaining the expressive information present in the pre-trained vectors. Further details on these are provided in Section 4.4.1 and Section 4.4.2.

Last but not the least, we wish to highlight that one crucial advantage of adopting this transformation based technique is that it provides the proposed model an added flexibility of acting as a "plug-and-play" module for other downstream tasks. Because the proposed approach does not jeopardize the word embedding training process, it allows end-users the liberty of choosing their own method of generating word embeddings and then utilize the proposed model as a means of post-processing step to gain interpretability.

4.4 Methodology

Our methodology section is divided into two sections. Section 4.4.1 describes the technique to infer the categorical representations. Then, Section 4.4.2 presents the details on how the transformation matrix is learned, and further discusses on how it induces the interpretability for word embeddings.

4.4.1 Inferring Categorical Embeddings

To infer the embeddings of categories, we leverage upon the dictionary definitions provided by the subject matter experts [100]. As an illustration, consider the definition of category "Disease or Syndrome" shown below:

Disease or Syndrome: *"A condition which alters or interferes with a normal process, state, or activity of an organism. It is usually characterized by the abnormal functioning of one or more of the host's systems, parts, or organs. Included here is a complex of symptoms descriptive of a disorder. Any specific disease or syndrome that is modified by such modifiers as acute, prolonged, etc. will also be assigned to this type. If an anatomic*

abnormality has a pathologic manifestation, then it will be given this type as well as a type from the Anatomical Abnormality hierarchy, e.g., Diabetic Cataract".

As these definitions are very precise, we leverage this expert knowledge and aim to infer the representation of "Disease or Syndrome". To do so, we first extract the medical concepts from their definitions and then use their already available word representations to infer their categorical meaning. Note that these medical concepts (underlined in the above example definition) are also present in our input pre-trained embeddings. Now, as the number of concepts contained in these categorical definitions is limited, this inevitably leads to a coarser estimation of their categorical meaning. To overcome this issue, we expand the set of associated medical concepts based on the external knowledge graph present in the bio-medical domain (the effectiveness of incorporating the neighbourhood set is validated in the experimental section). In this knowledge graph, the medical concepts are arranged in the form of a hierarchical tree (i.e., IS-A relationships). As such, the distance between the concepts in this tree denotes their semantic proximity. Building upon this premise, we assume that the concepts closer to each other in the hierarchy share greater information and thus the subtle cues obtained from the local neighborhood of concepts present in dictionary definitions might improve the overall categorical representation.

Formally, let $\mathbf{C} \in \mathbb{R}^{d \times m}$ denote the overall collection of categorical embeddings, d denote the embedding dimension and m denote the number of semantic categories. Now, to incorporate the above discussed local neighborhood information for concepts present in the dictionary definitions, a simple graph based scenario is considered. In this graph, nodes refer to the set of medical concepts and an edge is formed between concepts, if there is an hypernyms/hyponyms relationship between them. Let $\mathbf{V}_i = \{\mathbf{v}_{i1}, .., \mathbf{v}_{ij}\}$ denote the set of embedding vectors for medical concepts contained in the definition of i-th categorical concept, and $Neigh(\mathbf{v}_{ij})$ denote the corresponding set of local neighbours (siblings, parents and children) for the medical concept v_{ij}.

Our objective now is to infer the set of categorical embeddings $\hat{\mathbf{C}} = [\hat{\mathbf{c}}_i, ..., \hat{\mathbf{c}}_m]$ such that the categorical vectors are both close to the concepts present in their dictionary definitions and also to the local neighbours of the dictionary concepts. To achieve this, we propose the following loss function to infer their categorical representations:

$$L_c = \sum_{i=1}^{m} \left[\sum_{j=1}^{J} (||\hat{\mathbf{c}}_i - \mathbf{v}_{ij}||_2^2 + \sum_{k \in Neigh(\mathbf{v}_{ij})} \alpha ||\hat{\mathbf{c}}_i - \mathbf{v}_{ijk}||_2^2) \right] \tag{4.1}$$

where J denotes the number of dictionary concepts present in the particular category definition, and $\mathbf{v}_{ij}$, $\mathbf{v}_{ijk}$ represents the embeddings of dictionary concepts. The value of α is empirically set to 0.1 and is used to control the relative strengths between the concepts explicitly present in the dictionary definitions and their local neighbours. As it can be observed, the formulation is convex and its solution can be found by solving a system of linear equations. We minimize the categorical loss function and infer the categorical embeddings as follows:

$$\hat{\mathbf{C}} =_{\hat{\mathbf{C}}} L_c \tag{4.2}$$

The entire set of categorical embeddings is denoted as $\hat{\mathbf{C}} = [\hat{\mathbf{c}}_1, ..., \hat{\mathbf{c}}_m]$, and the closed form solution for $\hat{\mathbf{c}}_i$ is shown below:

$$\hat{\mathbf{c}}_i = \frac{\sum_{j=1}^{J} (\mathbf{v}_{ij} + \alpha \sum_{k \in Neigh(\mathbf{v}_{ij})} \mathbf{v}_{ijk})}{J + \alpha \sum_{j=1}^{J} K_{ij}} \tag{4.3}$$

where K_{ij} represents the size of $Neigh(\mathbf{v}_{ij})$.

4.4.2 Learning Transformation

To be precise, we expect our transformation technique to meet the following two objective: *1) the implicit conceptual property within the individual dimensions should be enlightened* and *2) the transformation should be carried out in such a way that the resultant embeddings retain the information present in the original vectors.* To accomplish the first goal, the idea is to attain a target space (after performing transformation) with the basis as the semantics of inferred categorical representations (refer Section 4.4.1). The corresponding value on individual dimension quantifies the amount of conceptual property being captured within these individual dimensions. Let $T : R^d \rightarrow R^m$ represent a linear transformation, and the transformed categorical embeddings are represented as $T(\hat{\mathbf{C}}) = [T(\hat{\mathbf{c}}_1), ..., T(\hat{\mathbf{c}}_m)]$. Since the transformed categorical embeddings act as a basis of the new space and these basis are also linearly independent unit vectors in the new space, an identity matrix could be used as a target for the transformed basis. To achieve this, we formulate the transformation as an optimization problem shown below:

$$\min_{\mathbf{W}} ||\mathbf{W}^T \cdot \hat{\mathbf{C}} - \mathbf{I}||_2^2 \tag{4.4}$$

Here the transformation matrix is denoted as $\mathbf{W}$ and $\mathbf{I}$ refers to an identity matrix. Note that this step acts as a soft regularization for linear independence, as in the real-word scenario, the distinct categorical embeddings may not be strictly independent of each other. In essence, this particular step of categorical basis conversion plays a vital role in inducing the interpretability in word vectors, and also allows us to explicitly define the meaning of the individual dimensions with their categorical types; thereby, enabling us to achieve our objective of performing dimension-wise interpretability.

Next, to meet our second objective of preserving the expressive features, we propose to regularize the transformation matrix by an orthogonal constraint. This is because of the peculiar property of orthogonal transformation to preserve the bilinear form i.e., Euclidean distance and Cosine in the latent space [95]. Since our transformation is from the original embedding space to an interpretable space, this may result in the change in number of dimensions; thereby, causing a possible information loss. To handle this, we adopt the principles of orthogonal transformation and mould that into our proposed optimization framework. This allows us to preserve the information particularly relevant to the categorical dimensions. The proposed orthogonal constraint is shown below:

$$\min_{\mathbf{W}} ||\mathbf{W}^T \cdot \mathbf{W} - \mathbf{I}||_2^2 \tag{4.5}$$

Now, since the focus of this study is to find a transformation matrix $\mathbf{W} \in R^{d \times m}$ that transforms the original pre-trained embeddings from d dimensional space to m dimensional space (that has inferred categorical embeddings as the basis), and the corresponding transformation matrix also attempts to preserve the information, the final objective to be minimized becomes the combination of these two objectives:

$$L_w = ||\mathbf{W}^T \cdot \hat{\mathbf{C}} - \mathbf{I}||_2^2 + \beta||\mathbf{W}^T \cdot \mathbf{W} - \mathbf{I}||_2^2 \tag{4.6}$$

Here β (empirically set to 0.2) controls the relative strengths of associations. To solve this, we take the gradient of our objective function (Equation 4.6) with respect to each of the model parameters and then adopt stochastic gradient descent to update our transformation matrix W:

$$\mathbf{W} \quad \leftarrow \quad \mathbf{W} - \eta \, \frac{\partial L_w}{\partial \mathbf{W}} \tag{4.7}$$

where η (empirically set to 0.001) is the learning rate for gradient descent. Overall, the fulfillment of two above discussed objectives allows us to achieve our goal of inducing the

interpretability in vector representation and concurrently retain the original expressive features.

4.5 Experiments

The focus of this section is to demonstrate the efficacy of the proposed model in improving the interpretability of the pre-trained word embeddings. In doing so, we first need a set of word embeddings trained on a massive corpora. For this purpose, we choose MEDLINE[1] - the largest available bibliographic repository in the domain of biomedicine. At this time of writing, it contains more than 24 millions records (articles) primarily from the research area of life sciences and biomedicine. Every article in MEDLINE is tagged with a set of special keywords known as Medical Subject Headings (MeSH) terms. Because they are assigned by subject-matter-experts, they find their utility in a variety of biomedical tasks. Thus, we believe that the use of MeSH terms (and correspondingly release of interpretable MeSH embeddings[2]) will have immediate practical benefits to the community.

Based on the full-scale MEDLINE corpus (and correspondingly MeSH terms), we use CBOW [16] word embedding model to train our embeddings. Additionally, as means of an alternate baseline, we also train another prominent word embedding model, namely, GloVe [17] on the same MEDLINE corpora. As suggested by the previous studies [18, 17], the number of embedding dimension is set to 300. Also, note the total number of semantic types (m) available is 133 [100].

4.5.1 Interpretability

(1) Qualitative Assessment of Interpretability

To perform the qualitative assessment of our results, we borrow experimental settings from the interpretable word embeddings literature [95, 93]. Specifically, the idea in qualitative evaluation is that if a particular vector dimension is interpretable then the top ranking words (from the entire vocabulary) for that dimension should display a form of semantic coherence. To examine this, we select four examples of biomedical significance [22, 23]. The selected examples are the following: a) Diabetes mellitus, b)

[1] https://www.nlm.nih.gov/pubs/factsheets/medline.html
[2] `https://github.com/kishlayjha/InterpretableMedicalEmbeddings`

Migraine disorders, c) Alzheimer's disease and d) Insulin. For each of these examples, we examine their top participating dimension and then look up for the top words with highest value in the same dimension. Table 4.1 presents the results for both pre-trained word embedding models (both CBOW and Glove) and the proposed model. As it can be observed, the semantic grouping of words resulted by CBOW/Glove is more or less arbitrary. In contrast, the results obtained by our transformed embeddings yields a meaningful semantic coherence. As an illustration consider the case of "Diabetes mellitus". For the proposed model, it can be observed (refer Table 4.1) that most of the terms in the group are closely related to the various aspects of "Diabetes" itself and the remaining few are related to the concept of "Disease" in general. In our transformed embeddings, we find the category name of these terms to be "Disease or Syndrome". Recall that as our transformation matrix is augmented with the categorical information, every dimension in the transformed vector is regularized by a particular categorical concept.

Another point we wish to highlight is the ability of the proposed model to answer question like: "To what extent a medical concept (e.g., Insulin) encodes the semantics of category *Pharmacological substance* or a *Hormone* within their dense dimensions". Note that the transformed embeddings have numerical values in their individual dimensions. These numerical values precisely help us in answering such aforementioned kind of questions. As an illustration, consider the case of "Insulin". In the medical domain, "Insulin" acts both as a pharmacological substance and a hormone. In our results, we obtained highest score for insulin in the category name - "pharmacological substance" and a relatively higher score in the category name - "hormone". From this result, one can speculate that the vector representations (generated by the state-of-the-art embedding models) of insulin captures the conceptual property of being a "pharmacological substance" more than that of a "hormone".

In essence, from the above discussed qualitative assessment it can be deduced that the proposed model is able to elucidate the meaning of individual dimensions and potentially shed insights into the notion of conceptual properties being captured by the state-of-the-art embedding models too. While informative, this form of qualitative assessment still does not inform us about the overall quality of the result set. To this end, a quantitative evaluation has to be performed.

(2) Quantitative Assessment of Interpretable Embeddings

Table 4.1: Qualitative evaluation of the original and generated embeddings

Concepts	CBOW	Glove	Proposed
Diabetes mellitus	25-hydroxyvitamin d 2, 3-hydroxyacyl coa dehydrogenases, whiplash injuries, youth sports, abdominal fat	humans, xanthomatosis, cerebrotendinous, glycogen, yang deficiency	diabetes insipidus, diabetes complications, diet therapy, digestive system diseases
Indomethacin	acute kidney injury, acute disease, "administration, oral", "abortion, septic", acidosis	acetohexamide, "administration, intravenous", agglutination, albumins, "4-aminobenzoic acid"	endothelin-1, endothelins, endothelin-1, endotoxemia, "endothelin-converting enzymes"
Alzheimer Disease	ac133 antigen, acinar cells, ablation techniques, abducens nerve diseases, acinar cells	"active transport, cell nucleus", "acid sensing ion channels", "abducens nerve diseases", "acinar cells", "actins"	amyotrophic lateral sclerosis, amyloidosis, "amyloidosis, familial", amyloid neuropathies, "amyloid neuropathies, familial"
Insulin	alpha-msh, artemia, anabolic agents, antithyroid agents, appetite	appetite, acromegaly, adrenalectomy, anabolic agents, andropause	insulin antagonists, insulin-like growth factor binding protein 1, insulin-like growth factor binding protein 2, lactation, lactation disorders

In order to perform a quantitative assessment, we analyze our results on a task much akin to semantic categorization. In more detail, every medical concept present in our vocabulary belongs to a certain number semantic categories. For instance, the medical concept "Diabetes mellitus" belongs to the semantic category of "Disease or syndrome". In this manner, every concept present in the dictionary is assigned a semantic category from the range of one to five. We probe whether the dimension with highest score (i.e., semantic labels predicted by proposed model) match the true semantic labels or not. Table 4.2 reports the accuracy for our Top-K dimensions. Now, as the previous studies do not perform dimension-wise interpretability, a direct comparison with their approach cannot be performed. For the sake of comparison, we developed a baseline

Table 4.2: Quantitative evaluation of semantic categorization task

Baseline	Accuracy (K=5)	Accuracy (K=10)	Accuracy (K=15)
Supervised	0.732	0.857	0.925
Proposed model (without neighbours)	0.423	0.557	0.652
Proposed Model	0.522	0.683	0.762

Table 4.3: Absolute values of correlation of the five measures relative to human judgments - MeSH-1

Measure	Physician	Expert
CBOW	0.8174	0.7632
GLoVe	0.8057	0.7541
Proposed model	0.8015	0.74328

(i.e., Supervised) that uses all the explicit semantic labels to train a linear model (using pre-trained emebddings) and reported the results in Table 4.2. As it can be observed, the proposed model (though unsupervised in nature) still maintains a reasonable performance as compared to the supervised model. Note that in our proposed model we do not use any explicit semantic labels. Now, in order to explore the effectiveness of incorporating the neighbour sets of medical concepts from the knowledge graph (refer Section 4.4.1), we evaluate the proposed model (with/without neighbourhood set) and report results. As it can be observed, the proposed technique of categorical inference significantly outperforms the baseline. We believe this is due to the ability of the proposed technique to obtain subtle cues from the informative neighbours of the dictionary concepts that ultimately improves the quality of categorical representation.

In summary, from Section 4.5.1, we can conclude that the proposed model has the capability to generate interpretable embeddings that have high proximity to the human intuition. While this accomplishes our core objective, we also aim to ensure that the information present in original pre-trained word vectors is retained in the transformed embeddings. To evaluate this, in Section 4.5.2, we report and analyze our results on the biomedical concept similarity/relatedness tasks.

Table 4.4: Absolute values of correlation of the five measures relative to human judgments- MeSH-2

Measure	Human expert
CBOW	0.7677
GLoVe	0.7586
Proposed model	0.7789

4.5.2 Expressive Performance

In this section, we inspect the expressive performance of our transformed embeddings as compared to the original vectors.

(1) Evaluation Datasets

To examine the ability of transformed embeddings to retain original information, we choose biomedical concept similarity/relatedness task. The evaluation sets (i.e., MeSH-1 and MeSH-2) are borrowed from [42] and [43] respectively. Both datasets consist of 30 and 36 concept pairs that were manually rated by human experts indicating their semantic similarity.

(2) Results and Discussion

MeSH-1

Table 4.3 presents the Spearman (ρ) coefficient values obtained after applying the proposed model on the first dataset (MeSH-1). As it can be observed from the table, the proposed model performs on par with both CBOW and GloVE and achieves similar correlation as pre-trained embeddings with both physician's and experts judgments.

From the results, it can be inferred that the transformed embeddings retain the features of original vectors. We believe the reason for this lies in the orthogonality constraint imposed on the learned transformation. Because such form of imposition leverages the principles of orthogonal transformation (the has unique capability of preserving the bilinear form), the categorical related information loss is minimized.

MeSH-2

Table 4.4 shows the correlation values obtained for the Spearman (ρ) coefficient for MeSH-2 dataset. Note that in this dataset, the proposed model obtains even higher

correlation value as compared to the state-of-the word embedding models. Analyzing this result further, we believe that the reason for this lies in the capability of the proposed model to preserve the relevant information related to categorical dimensions in the transformed space, and correspondingly removing the unrelated information.

4.6 Conclusions

In this study, we proposed a novel framework that induces the interpretability of word embeddings in the medical domain. Specifically, by leveraging upon the principles of category theory and rich categorical knowledge present in the biomedical domain, the model learns a transformation matrix that induces the interpretability of word embedding dimensions at a granular level. The transformation matrix in particular is learned in a such a way that any pre-trained input embeddings can be transformed to a new space where the produced embeddings reveal the conceptual meaning hidden within their individual dimensions and concurrently posses the expressive features present in the original pre-trained vectors.

Part II

Building a Continual Representation Learning Model

Chapter 5

Learning Continual Representations for Bipartite Networks

5.1 Introduction

Bipartite networks are a special class of complex networks that contain two distinct types of entities, and the relational ties only exist between entities (or nodes) of different types. Many real-world biomedical networks pertain to a native bipartite structure, where one class of the nodes is usually comprised of cellular components such as genes, miRNAs or proteins, and the other class is composed of various indicators of human diseases such as symptoms and drug effects. Effective data analysis on this ubiquitous network structure can benefit a multitude of practical applications, such as identifying casual pathways in gene-phenotype networks, predicting new targets for existing drugs in drug-target networks, and assisting clinical decision making via patient-symptom graphs [101]. However, traditional network analysis methods suffer from high computational and space costs. To overcome this challenge, network representation learning, an area of research that aims to learn low-dimensional vector representations of nodes has attracted significant attention. These node representations (or embeddings) are learned such that the connectivity structure of the network is preserved in the learned vector space. Recent literature has intensively studied this topic and various approaches ranging from matrix factorization [102], random walk based [103] to graph convolution [104]

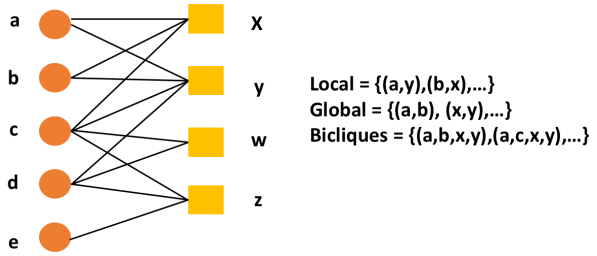


Figure 5.1: An example of a bipartite network with various topological properties.

have been proposed. While a majority of these approaches have been developed for homogeneous/heterogeneous networks, some recent studies such as [105] have attempted to model the special properties of bipartite networks. Despite significant advances made, the existing approaches still have certain inherent drawbacks. First, the approaches fail to model the unique high-order structures (e.g, bicliques) present in the bipartite networks. This is limiting because bicliques are the smallest high-order structures that characterize the bipartite networks. More importantly, bicliques facilitate a principled approach to analyze the biomedical bipartite networks as they are capable of addressing critical biological challenges in the biomedical applications such as biclustering microarray data, identifying common gene-set associations, and integrating diverse functional genomics data [106, 107]. Second, the existing approaches mainly assume a static setting for networks. However, real-world biomedical networks are continually evolving. To address these issues, we propose a new representation learning approach that accurately preserves the topological properties of bipartite networks, and at the same time efficiently updates the node representations in consecutive network snapshots. To accurately preserve the intricate bipartite structure, we design a customized autoencoder that maximally reconstructs the structural proximity between nodes in the learned embedding space. Figure 5.1 presents an illustrative example of a bipartite network with various structural components. Here, the vertexes of distinct node types that are linked by a direct edge (e.g. [a,y]) characterizes the local structure. In contrast, the vertices that share common neighbors (e.g. [a,b]) but have no direct link constitute the global structure. Finally, the vertices (e.g. [a,b,x,y]) that participate in the 2×2 sub-graph represents the biclique structure. To effectively encode these structural properties into node representations, we design a dedicated objective function for each component and then propose a joint inference mechanism.

Meanwhile, to efficiently generate the representations in a dynamic setting, we design an incremental learning strategy that interleaves the proposed structure-preserving

technique with the central principles of continual machine learning [108]. Specifically, the approach considers the successive network snapshots as a sequence of related tasks and carefully updates the node representations affected by the new snapshot, while preserving the representations that are well-trained previously. The main challenge in this strategy is to automatically identify the parameters that are subject to retraining and retention. To address this, we propose the following: at every new network snapshot, we quantify the importance of parameters according to their contribution to the loss function. Then, the important parameters are frozen to preserve the current knowledge and the remaining are used for future training. This process is iteratively applied to the consecutive snapshots and the node representations are obtained promptly.

In this research, our contributions can be summarized as:

- We propose a new representation learning approach that is tailored for bipartite networks. Notably, this class of network has special usability and implications in the field of network biology and medicine.
- The proposed structure-preserving technique identifies and models the intricate topological properties (i.e., local, global and biclique) such that the unique bipartite structure is accurately preserved.
- We propose a continual learning scheme that updates the representations in an online fashion. This strategy greatly improves the computational efficiency of proposed approach whilst accounting for the rapidly evolving nature of the biomedical bipartite networks.
- Extensive experiments on real-world biomedical datasets through the tasks of network reconstruction, link prediction, and recommendation validates the effectiveness of the proposed approach.

5.2 Related Work

5.2.1 Network Embedding

For a recent survey on this topic, please refer [109]. The initial NRL approaches mainly used matrix factorization (MF) based techniques. However, the traditional MF approaches suffered from scalability issues. To mitigate this, recent NRL research leveraged upon the advances in deep neural networks and developed powerful approaches

such as DeepWalk [110], LINE [111], node2vec [103], and SDNE [112]. Building upon this research, the authors in [113] proposed an approach named DynGEM that learns the node representations in dynamic networks. More recently, the authors in [105] proposed an approach named BiNE that models the vertex type information of bipartite networks. However, we differ from them in two aspects. First, BiNE misses to model the unique high-order structure (i.e., bicliques) present in the bipartite networks. Second, BiNE is designed for static networks, and thus is unable to obtain the representations in a dynamic setting.

5.2.2 Network Embedding In Biomedicine

While the network embedding approaches have been widely evaluated on social and information networks, their investigation with biomedical networks is recent. In biomedicine, network embedding techniques have been broadly applied to pharmaceutical data analysis, multi-omics data analysis, and clinical data analysis [101, 114, 115]. Studies such as [101] introduced DeepWalk [110] to learn the concept representations in heterogeneous biological knowledge graphs. In another study [114], the authors integrated NLP techniques with network embedding, and demonstrated significant improvement in the task of drug-disease interaction. While these approaches elucidated the practical benefits of NRL in biomedicine, they did not factor in the evolving nature of biomedical data. The proposed approaches continually acquires new information, and correspondingly updates the node representations over longer time-spans.

5.2.3 Continual Machine Learning

Continual learning [108, 116] is an area of research that is useful when the data arrives in streams or snapshots. Prior research has studied continual learning in the context of supervised, semi-supervised, and unsupervised learning. While our research builds upon the ideas of continual learning, it differs from the existing approaches in two aspects. Firstly, the goals are different. Unlike the existing studies that mainly focus on multi-task learning, the proposed approach is designed for single-task incremental settings. Second, the existing approaches are proposed for computer vision focused embedding models that mainly utilize the imaging datasets. However, the imaging datasets do not share the topological properties of the bipartite network datasets. As a consequence, the computer vision focused approaches cannot be directly applied to the current problem

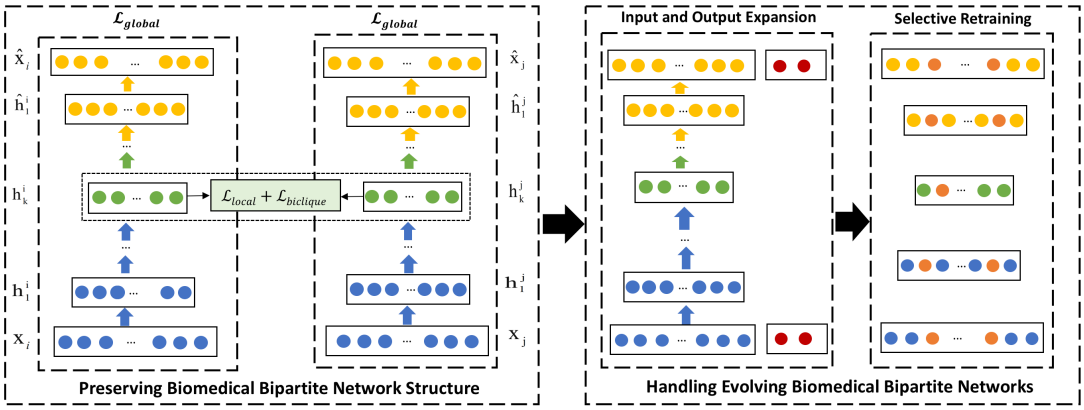


Figure 5.2: Continual representation learning framework for bipartite networks: The figure (left) shows the deep autoencoder model that preserves the intricate bipartite structure from three perspectives (i.e., global, biclique and local). The figure (right) shows the input/output expansion and selective retraining mechanisms to update the representations in an online fashion.

setting.

5.3 Approach

5.3.1 Problem Formulation

Let $G = (U, V, E)$ denote a bipartite network, where $U = \{u_1, ..., u_n\}$ and $V = \{v_1, ..., v_m\}$ are the two sets of distinct vertices, and $E \subseteq U \times V$ defines the set of links between them. Each edge e_{ij} is associated with a weight that denotes the strength of relationship between connected vertices u_i and v_j. The weighted adjacency matrix for G is denoted as $\boldsymbol{A} \in \mathbb{R}^{n \times m}$. If $e_{ij} \in E$, $a_{ij} > 0$; otherwise, $a_{ij} = 0$. The i-th row of the adjacency matrix is denoted as $\boldsymbol{A}_i = [a_{i1}, ..., a_{im}]$. Now, given a series of network snapshots, i.e., $G = \{G_1, ..., G_T\}$, where $G_t = (U_t, V_t, E_t)$ and T is the number of snapshots, the continual representation learning aims to learn a time-series of mappings $F = \{f_1, ..., f_T\}$ such that the function f_t maps each node in G_t to a d-dimensional embedding space.

5.3.2 Overview of Proposed Model

A promising continual representation learning approach should be able to generate node embeddings such that the embeddings preserve the intricate bipartite structure and flexibly updates the representations to accommodate the newly available data. To achieve this, we design a customized autoencoder architecture that reconstructs the bipartite network structure from three perspectives: a) global structure (Section 5.3.3), b) biclique structure (Section 5.3.4), and c) local structure (Section 5.3.5). Unlike the existing bipartite embedding approach [105] that is unable to model the non-linear properties of networks, the proposed approach employs a multi-layer autoencoder architecture (consisting of non-linear functions) that maps the data into a highly non-linear latent space and effectively captures the non-linearity. Further, to continually accommodate the new data, the proposed autoencoder model is carefully retrained such that only the representations that are affected by the new network snapshot are retrained while the remaining are simply retained. Figure 7.2 shows an overview of the proposed model.

5.3.3 Modeling Global Structure

The global structure of a bipartite network is described by the similarity of node pairs neighborhood structure. Specifically, this structure attempts to model the implicit relations between vertices of the same type. For any node of type $u_i \in U$, its neighborhood structure is defined as $N(u_i)$. Then, the proximity between two nodes of the same type U is defined as:

$$S_{ij}^{U} = \frac{|N(u_i) \cap N(u_j)|}{\sqrt{d_i d_j}}, u_i, u_j \in U, \tag{5.1}$$

where $N(u_i) = \{v_j \in V | a_{ij} > 0, u_i \in U\}$ represents the neighbourhood set of node u_i. d_i and d_j refers to the degree of nodes u_i and u_j respectively. Similarly, for any node of type $v_i \in V$, the proximity between nodes of type V is:

$$S_{ij}^{V} = \frac{|N(v_i) \cap N(v_j)|}{\sqrt{d_i d_j}}, v_i, v_j \in V, \tag{5.2}$$

In this way, the implicit relation between nodes is obtained by the similarity matrix $\boldsymbol{S^U}$ (or $\boldsymbol{S^V}$). Meanwhile, the proximity relationship between vertices is also characterized in the adjacency matrix $\boldsymbol{A}$. Thus, to obtain a comprehensive representation of relationships between vertices, we introduce an extended matrix $\boldsymbol{A'} \in R^{(n+m)\times(n+m)}$.

$$\boldsymbol{A'} = \begin{bmatrix} \boldsymbol{S^U} & \vdots & \boldsymbol{A} \\ \cdots & \cdots & \cdots \\ \boldsymbol{A^T} & \vdots & \boldsymbol{S^V} \end{bmatrix}$$

Given the matrix $\boldsymbol{A}^{'}$, the global structure of a node $x_i \in \{U \cup V\}$ is represented by the vector $\boldsymbol{a}_i^{'}$ (i.e., $\boldsymbol{x_i}$). This global structure can be modeled by an autoencoder that consists of two parts: a) encoder and b) decoder. Both parts contain multiple layers of non-linear function that map the input data to the reconstruction space. Thus, given an input data $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{n+m}$ and reconstructed data $\hat{\boldsymbol{X}} = \{\hat{\boldsymbol{x}}_i\}_{i=1}^{n+m}$, the hidden representations for each layer in the encoding procedure is shown as follows:

$$\begin{aligned} \boldsymbol{h}_1^i &= f(\boldsymbol{W}_1 \boldsymbol{x}_i + \boldsymbol{b}_1), \\ \boldsymbol{h}_k^i &= f(\boldsymbol{W}_k \boldsymbol{h}_{k-1} + \boldsymbol{b}_k), k = 2, ..., K, \end{aligned} \tag{5.3}$$

where f denotes the sigmoid activation. $\boldsymbol{h}_k^i$ denotes the representation of the k-th hidden layer. $\boldsymbol{W}_k$ and $\boldsymbol{b}_k$ denote the k-th hidden layer's weight matrix and bias respectively. The calculation procedure is reversed for decoder and the hidden representations for each layer is calculated as:

$$\begin{aligned} \hat{\boldsymbol{h}}_{k-1}^{\ i} &= f(\hat{\boldsymbol{W}}_k \hat{\boldsymbol{x}}_i + \hat{\boldsymbol{b}}_k), k = K, ..., 2, \\ \hat{\boldsymbol{x}}_i &= f(\hat{\boldsymbol{W}}_1 \hat{\boldsymbol{h}}_1 + \hat{\boldsymbol{b}}_1), \end{aligned} \tag{5.4}$$

where $\hat{\boldsymbol{h}}_{k-1}^{\ i}, \hat{\boldsymbol{W}}_k$ and $\hat{\boldsymbol{b}}_k$ denote the hidden representations, weight matrix, and bias term of the reconstruction layer respectively. The loss function to minimize the reconstruction error is defined as follows:

$$\mathcal{L}_{glob} = ||(\hat{\boldsymbol{X}} - \boldsymbol{X}) \odot \boldsymbol{B}||_F^2 \tag{5.5}$$

In Equation 5.5, $\odot$ denotes the Hadamard product and $\boldsymbol{B}$ denotes the weight matrix. Each weight vector $\boldsymbol{b}_i = \{b_{ij}\}_{j=1}^{n+m}$ in matrix $\boldsymbol{B}$ is defined as:

$$b_{ij} = \begin{cases} \alpha > 1, & a_{ij}^{'} > 0 \\ 1 & a_{ij}^{'} = 0 \end{cases} \tag{5.6}$$

where $a_{ij}^{'}$ is the j-th elements of $\boldsymbol{a}_i^{'}$ and α is the hyper-parameter. The weight matrix $\boldsymbol{B}$ is introduced to enforce greater penalty to the reconstruction loss of non-zero elements than that of zero elements. Overall, the proposed approach exploits the power of deep autoencoder to preserve the neighborhood structure, thus making them robust to the sparse nature of biomedical bipartite networks. Notably, the proposed approach is

different from the existing approaches such as BiNE [105] that perform biased random walk to characterize the global structure.

5.3.4 Modeling Bicliques

Bicliques are the subgraphs that characterize the smallest cohesive structure in bipartite networks [117]. While modeling the neighborhood structure captures the global structure, it is important to encode such high-order structures to preserve the overall structure accurately. To encode these structural units into feature representations, we first utilize the biclique algorithm [118] to enumerate bicliques and then perform biclique expansion to identify the implicit links. For instance, the expansion of biclique shown in Figure 5.1 is the following: clique_pairs = {(a,b), (x,y), (a,c), (x,y)}. Following the expansion, we create a clique matrix $\mathbf{C}$, where each element c_{ij} is defined as the following:

$$c_{ij} = \begin{cases} 1 & \text{if pair } (i,j) \text{ is present in clique\_pairs} \\ 0 & \text{otherwise} \end{cases} \tag{5.7}$$

Since the vertices participating in bicliques form a cohesive structure, we speculate that retaining the proximity of participant nodes in the embedding space will enable us to retain the network structure more accurately. The loss function for capturing this relationship can be formulated as:

$$\mathcal{L}_{biclique} = \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}||\mathbf{h}_z^i - \mathbf{h}_z^j||_2^2 \tag{5.8}$$

where $\boldsymbol{h}_z^i$ and $\boldsymbol{h}_z^j$ denote the representations of the output layer.

5.3.5 Modeling Local Structure

A direct edge between vertices of different types in a bipartite network provides the explicit structure information. For instance, in a citation network, if an article cites another then they should share some common topic. Modeling this type of explicit relation enables us to capture the local structure of the network. The loss function for capturing the local structure of a bipartite network can be formulated as:

$$\mathcal{L}_{local} = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}||\mathbf{h}_z^i - \mathbf{h}_z^j||_2^2 \tag{5.9}$$

Minimizing the loss function $\mathcal{L}_{local}$ makes two nodes with direct links to be mapped close in the embedding space, thus preserving the local network structure. Different

from BiNE [105] that uses the inner product to model the interaction between two entities, we utilize the hidden representations of two directly connected nodes from two parallel deep autoencoders to preserve the explicit relations.

5.3.6 Joint Optimization

To embed a bipartite network by preserving global, biclique and local structural units simultaneously, we combine their objective functions to form a joint optimization framework:

$$\mathcal{L} = \mathcal{L}_{local} + \lambda_1 \mathcal{L}_{glob} + \lambda_2 \mathcal{L}_{biclique} + \lambda_3 \mathcal{L}_{reg} \tag{5.10}$$

where λ_1, λ_2 and λ_3 are the balancing parameters. $\mathcal{L}_{reg}$ is the regularization term that prevents overfitting, which is defined as follows:

$$\mathcal{L}_{reg} = \frac{1}{2} \sum_{k=1}^{K} (||\mathbf{W}^{(k)}||_F^2 + ||\hat{\mathbf{W}}^{(k)}||_F^2) \tag{5.11}$$

To optimize the joint model, we run the stochastic gradient descent algorithm until convergence.

5.3.7 Generalizing To Evolving Bipartite Networks

In this section, we describe our efforts to handle the evolving nature of biomedical bipartite networks. Given a stream of network data for T snapshots, $t = 1, ..., t, ..., T$, the objective is to efficiently update the node representations in successive network snapshots (G_t). To address this, we propose an incremental learning strategy that expands and selectively retrains [116] the autoencoder to fit the data distribution of the incoming network snapshot. Since the proposed approach exploits the power of deep neural networks, selective retraining can be done in a straightforward manner through retraining of the learned network weights. Once the selective retraining is complete, we perform temporal alignment over the representations to ensure that the embeddings evolve smoothly. Overall, the proposed incremental learning strategy allows the model to scale to larger networks without compromising the prediction accuracy.

Formally, at time t, the proposed incremental strategy aims to learn the autoencoder parameters $\boldsymbol{W}^t$ by solving the following problem:

$$\min_{\boldsymbol{W}^t} \mathcal{L}(\boldsymbol{W}^t; \boldsymbol{W}^{t-1}, G_t) \text{ , t=1,...,T} \tag{5.12}$$

where $\mathcal{L}$ is the loss function defined in Equation 5.10. $\boldsymbol{W}^{t-1}$ is the representations learned at $t-1$ and acts as a prior knowledge. This formulation enables an effective

mechanism to learn and transfer useful knowledge among the network snapshots. Now, with each new network snapshot, the number of vertices may vary. To address this issue, we propose the following: if the number of vertices at time t is smaller than the number of vertices at time $t-1$, we add the corresponding number of isolated vertices to the network. On the contrary, if the number of vertices at time t is greater than the number of vertices at time $t-1$, the number of input and output neurons of the autoencoder is increased to meet the size of input data (for illustration refer red nodes in Figure 7.2). The change in the size of input induces two parameter matrices expansion: $\boldsymbol{W}_1^t = [\boldsymbol{W}_1^{t-1}, \boldsymbol{W}_1^L]$ and $\boldsymbol{W}_z^t = [\boldsymbol{W}_z^{t-1}, \boldsymbol{W}_z^L]$, where $\boldsymbol{W}^L$ denotes the expanded weight matrix and z is the output layer. Once the input and output layer of the autoencoder is expanded, the model can adapt to the new shape of data. This mechanism allows us to pre-train the newly added weights and preserve the weights that are well-trained previously. The optimization formulation is shown below:

$$\min_{\boldsymbol{W}_k^L} \mathcal{L}(\boldsymbol{W}_k^L; \boldsymbol{W}_k^{t-1}, G_t) \text{ , k=1,z} \tag{5.13}$$

Having adapted to the new shape of data, the next step is to update the node representations without retraining the model from scratch. To accomplish this, we propose to selectively retrain the network parameters (for illustration refer orange nodes in Figure 7.2). More specifically, the idea is to explicitly retrain the weights that are affected by the new network snapshot and retain the other weights that are well-trained previously. To achieve this, we filter the neurons based on their contribution to the loss function. The contribution score is developed from the Taylor expansion of the loss function. Basically, it represents the difference between the loss with and without each neuron. In other words, if the removal of a neuron leads to relatively small accuracy degradation, then this unit is recognized as an unimportant unit and vice-versa. This form of local sensitivity based ranking strategy effectively factors in the background information while computing the informativeness of concepts/samples for performing dynamic updates. Different from techniques that adopt global sensitivities, i.e., selecting concepts at a topic-level, the proposed approach accounts for neighborhood semantics at a granular level [119]. Technically, the contribution of a parameter can be quantified by the error induced after removing it. The induced error can be measured by squared difference of prediction errors with/without the parameter w_m.

$$I_m = \mathcal{L}(\boldsymbol{W}^t; \boldsymbol{W}^{t-1}, G_t) - (\mathcal{L}(\boldsymbol{W}^t | w_m^t = 0); \boldsymbol{W}^{t-1}, G_t)^2 \tag{5.14}$$

As computing I_m for each parameter w_m^t is computationally expensive, we approximate

I_m in the vicinity of $\boldsymbol{W}^t$ by its first-order Taylor expansion which simplifies to:

$$I_m^{(1)}(\boldsymbol{W}^t) = (g_m^t w_m^t)^2 \tag{5.15}$$

where $g_m^t = \frac{\partial \mathcal{L}}{\partial w_m^t}$ are the elements of gradient $\boldsymbol{g}$. Based on this contribution score, we sort the neurons unit layer by layer and identify top β units. We consolidate the units that contribute little to the final loss into a sub-network S. Then, we utilize them for retraining in the next snapshot by solving the following problem:

$$\min_{\boldsymbol{W}_S^t} \mathcal{L}(\boldsymbol{W}_S^t; \boldsymbol{W}_S^{t-1}, G_t) \tag{5.16}$$

where $\boldsymbol{W}_s$ are the weights of the selected subnetwork S. In this way, the selective retraining is accomplished for every consecutive snapshot. Finally, to ensure the stability of embeddings over successive snapshots and prevent the catastrophic forgetting, we perform temporal alignment [67].

5.4 Experiments

In this section, we conduct experiments and analysis on the publicly available biomedical bipartite networks. Below, we describe the chosen datasets and Table 5.1 reports their overall statistics.

- Biological General Repository for Interaction Datasets (BioGRID) [120]: BioGRID is a publicly available bipartite interaction network consisting of two types of nodes, where the nodes represent gene and protein respectively, and the edge weight indicates the strength of relationship between them.

- PubTator [121]: This is a bipartite network dataset created from articles present in PubMed [3]. The nodes in this dataset contain diseases and genes, where the edge weight represents the co-occurrence of gene-disease in the same article.

- Disease-Symptom [3]: This is a bipartite dataset that depicts relationship between diseases and symptoms. The edge weight refers to the co-occurrence of disease-symptom in the same article.

5.4.1 Baselines

The following benchmark network embedding algorithms are chosen to examine the performance of the learned representations. Since the majority of baseline algorithms

Table 5.1: Statistics of the chosen biomedical datasets

	—V—	—E—	T
BioGRID	1,000-1,268	65,000-105,195	70
PubTator	3,510-4,523	143,190-322,590	110
Disease-Symptom	6,123-9,126	66,435-171,025	120

are designed for static networks, we apply them independently to each snapshot and then rotate the embeddings as in [67] for alignment.

- Deepwalk [110]: This algorithm learns representations by leveraging skip-gram with truncated random walk technique.
- LINE [111]: LINE learns node representations by optimizing both first-order and second-order proximity.
- Node2Vec [103]: Node2Vec designs a biased random walk to generate a corpus of node sequences, and then adopts the strategy similar to DeepWalk to generate representations.
- SDNE [112]: SDNE is an autoencoder based model that learns representations by capturing the non-linearity of networks.
- Metapath2Vec [122]: Metapath2Vec is a heterogeneous network embedding algorithm that formalizes meta-path based random walks to construct the heterogeneous neighborhood of a node and then adopts skip-gram model to produce representations.
- DynGEM [113]: DynGEM is a dynamic network embedding algorithm that employs deep autoencoder at its core and generates stable embeddings over time.
- BiNE [105]: BiNE is a recent bipartite network embedding approach that models the explicit and implicit relations simultaneously by performing a biased and self-adaptive random walk.

5.4.2 Results and Discussion

In our experiments, we evaluate the performance of the proposed approach and baseline algorithms on the tasks of network reconstruction, link prediction, and recommendation.

Table 5.2: Network reconstruction performance on biomedical datasets

Models	BioGRID	PubTator	Disease-Symptom
DeepWalk	0.325	0.314	0.098
LINE	0.452	0.431	0.119
Node2Vec	0.553	0.514	0.214
SDNE	0.801	0.697	0.594
Metapath2Vec	0.803	0.701	0.608
DynGEM	0.812	0.703	0.618
BiNE	0.821	0.729	0.635
Proposed	**0.861**	**0.753**	**0.674**

Table 5.3: Link prediction performance on biomedical datasets

Models	BioGRID	PubTator	Disease-Symptom
DeepWalk	65.18	77.11	67.18
LINE	67.12	80.33	69.11
Node2Vec	69.13	82.11	71.08
SDNE	71.49	83.92	72.93
Metapath2Vec	75.34	85.11	75.12
DynGEM	79.34	87.11	78.11
BiNE	81.48	90.91	82.11
Proposed	**84.22**	**94.13**	**86.50**

Network Reconstruction

The objective of this experiment is to examine the capability of node representations generated by various approaches to accurately reconstruct the network. Specifically, we learn the node representations over various networks (i.e., BioGRID, PubTator and Disease-Symptom) and predict the links between pair of vertices in the corresponding networks. Since the existing links in the networks are already known, these can act as our ground-truth. The pairs of vertices are ranked according to their corresponding reconstructed proximity. Then, we calculate the ratio of real links in top-k pairs of

vertices as the reconstruction precision (i.e., the training set error, of different methods). Table 6.7 reports the Mean Average Precision (MAP) averaged over the entire network snapshot for each dataset. From the results, we can observe that the proposed method achieves significant improvement over the baselines in all the datasets. The results indicate that the proposed approach is able to reconstruct the network structure in an accurate manner. Among the baselines, the existing bipartite network embedding approach (BiNE) performs better than other homogeneous/heterogeneous embedding approaches. This demonstrates the importance of modeling special bipartite properties for learning quality node representations. Moreover, it emphasizes the necessity of developing approaches that are tailored for the bipartite networks. Analyzing the results further, we observe that both SDNE and DynGEM perform better than other contemporary homogeneous network embedding approaches. We speculate that the reason for this lies in the capabilities of both SDNE and DynGEM to capture the non-linearity of networks.

Link Prediction

To examine the performance in link prediction, we follow the experimental protocol proposed in BiNE [105]. Specifically, for all the datasets, the observed links are treated as positive instances, and an equal number of random (unobserved) node pairs are considered as the negative instances. For each of our datasets, we randomly selected 85% of the data as training, 5% for validation, and the remaining 10% as test. Each network embedding algorithm is trained on the training data and the node embeddings are generated. The node embeddings of connected node pairs are concatenated to obtain the representations of edges. These edge representations are treated as feature vectors for training a logistic regression classifier, and whether or not a node pair has edge (link) as the ground truth. The logistic regression classifier is trained on the training data, and its performance is evaluated on the test data. We use ROC curve (AUC-ROC) as the evaluation metric. Table 5.3 reports the results. The results demonstrate the capability of the proposed approach to capture the relevant links. Moreover, the gain in performance with respect to BiNE highlights the importance of incorporating bicliques that nudge the node representations to be robust.

Table 5.4: Recommendation performance on biomedical datasets

	BioGRID		PubTator		Disease-Symptom	
	F1@10	MAP@10	F1@10	MAP@10	F1@10	MAP@10
DeepWalk	5.82	4.11	4.28	6.32	8.50	9.67
LINE	9.62	8.94	7.81	9.12	8.99	11.11
Node2Vec	6.73	6.07	6.25	7.11	8.54	10.23
SDNE	10.33	9.65	15.67	17.19	15.38	18.28
Metapath2Vec	11.45	13.45	17.67	19.32	16.12	20.45
DynGEM	12.18	15.83	18.97	23.98	19.23	23.03
BiNE	17.45	22.98	21.57	27.32	22.54	24.11
Proposed	**20.21**	**25.89**	**23.18**	**31.67**	**25.81**	**27.76**

Table 5.5: Effect of local, global and biclique on network reconstruction

Models	BioGRID	PubTator	Disease-Symptom
Local	0.752	0.651	0.580
Global	0.801	0.702	0.621
Biclique	0.781	0.683	0.604
Local+Biclique	0.811	0.714	0.634
Local+Global	0.821	0.729	0.645
Global+Biclique	0.844	0.737	0.652
Proposed	**0.861**	**0.753**	**0.674**

Recommendation

Given a network, the objective is to estimate the preference of an entity u_i (e.g., gene) to associate with another entity v_j (e.g., disease). Similar to link prediction, we split 85% of the links in the datasets as training, 5% as validation, and the remaining links as test set. For a gene and a disease in the training set, we use the inner product of their embedding to evaluate the gene's binding preference for the disease, and for each gene, we select $n = 10$ items with a largest preference scores for recommendation. We run the experiment 10 times and average the performance. Both F1@10 and MAP@10 are reported in Table 5.4. In this particular task, we observe that homogeneous network embedding approaches have relatively poor performance. The reason for this lies in the inability of homogeneous approaches in modeling the disease-gene bipartite networks properties, which is especially important for recommendation tasks. This is also supported by the observation that BiNE perform significantly better than others in this task. Overall,

the proposed method outperforms baseline methods in all three data sets. This result reinforces the importance of modelling the fundamental structural units (i.e., global, bicliques, and local) of bipartite networks in a joint manner.

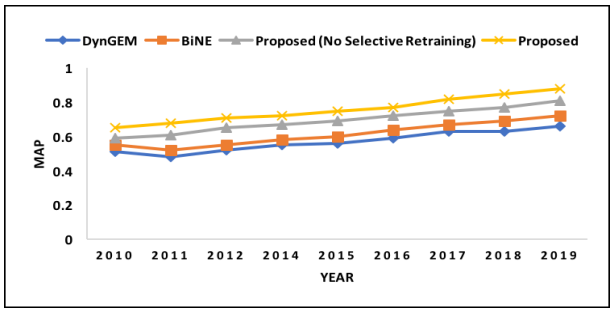


Figure 5.3: Mean Average Precision of various approaches on PubTator network snapshots.

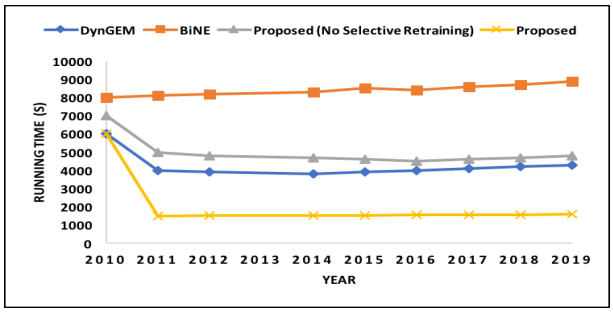


Figure 5.4: Runtime Performance of various approaches PubTator network snapshots.

Effect of Local, Global and Bicliques

To analyze the contribution of each topological units (i.e., local, global and bicliques), we develop multiple variants of the proposed model by removing individual components and generate the feature representations. Then, we evaluate the performance of representations on the task of network reconstruction. Table 5.5 reports MAP results over snapshots on all three datasets. As it can be seen from the table, each individual structural units contributes uniquely towards the structure of bipartite networks. We also observe that the performance of bicliques is greater than local. This is reasonable because bicliques are the intricate units that characterize the community structure of bipartite networks. Further, we note that the contribution of "Global+Biclique" is more

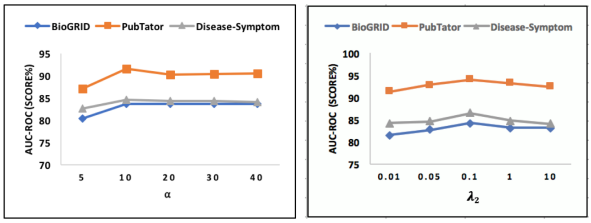


Figure 5.5: Impact of hyper-parameter values α and λ_2 on the task of link prediction.

than "Local+Global" or "Local+Biclique". This indicates the critical role of high-order structures in preserving the network structure accurately. In summary, the overall results validate the importance of incorporating various topological units in preserving the network structure in a more effective manner.

Effect of Continual Training on Computational Efficiency

In this section, we analyze the effect of continual training upon the computational efficiency of various embedding approaches. The comparison includes the time taken to compute embeddings and MAP performance of various approaches in the network snapshots (2010-2019) of the PubTator dataset. Figures 5.3 and 5.4 show the results. To compare the performance, we choose two baselines: a) BiNE and b) DynGEM. It can be observed that the proposed approach is faster than both BiNE and DynGEM in most of the network snapshots. We speculate this is due to the inability of BiNE to account for the temporal dynamics of biomedical bipartite networks. As compared to DynGEM that considers dynamic graphs, the proposed approach yield better performance due to the design choice of selective retraining. To better examine the benefits of selective retraining, we design a variant of model (Proposed - No Selective Retraining) that does not perform selective retraining. Instead, the model retrains from scratch at every incoming network snapshot. From the results, we observe that the computational efficiency of DynGEM is better than the variant model. However, the variant model performs better in terms of MAP. This is reasonable because DynGEM do not model the unique topological properties of bipartite networks. Overall, the results demonstrate that the design choice of selective retraining plays a critical role in improving both the fidelity and learning efficiency of the proposed approach.

5.4.3 Hyper-Parameter Settings

The experiments were carried out on NVIDIA TITAN Xp GPU. The proposed approach uses a 2-layer autoencoder. For a systematic comparison of results, we set the size of embeddings produced from both the proposed approach and the baseline algorithms to 100. The hyper-parameters of the loss functions are tuned by using grid search on the validation set. Following the convention in existing studies [105, 111], we set the range to [0.01, 0.05, 0.1, 1, 10]. The optimal hyper-parameters values are set to $\lambda_1 = 10$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.01$ respectively. We study the sensitivity of biclique parameter (λ_2) by fixing the others. Figure 5.5 (right side) shows the results. The best performance is obtained at 0.1 and then the performances starts to decrease. Similarly, we test the learning rate and the weight α of reconstruction loss for non-zero elements in the range of [0.01, 0.025, 0.1] and [5, 10, 20, 30, 40] respectively. The optimal values for learning rate and α are set to 0.025 and 10. From the Figure 5.5 (left side), we observe that introducing the parameter α is useful. The results initially improve when we increase the value of α, however, they stabilize at the higher values.

5.5 Conclusions

In this research, we proposed a new representation learning approach for the bipartite networks. The proposed approach designs a structure-preserving technique that models the unique topological properties of bipartite networks and preserves the intricate structure accurately. Moreover, the proposed approach develops a continual learning scheme that progressively acquires the newly available information, and adapts the representations to reflect the up-to-date knowledge. Extensive experiments conducted on the real-world biomedical networks validate the efficacy of the proposed approach, and suggests that the proposed framework is capable of generating meaningful representations that are useful for a variety of downstream biomedical applications.

Chapter 6

Knowledge-Guided Continual Representation Learning

6.1 Introduction

Around 70% of the total web search queries are of medical and healthcare category [123]. Consequently, there is a growing interest among practitioners to develop sophisticated text mining and natural language processing (NLP) systems that can handle the unique challenges posed by the biomedical domain. A precursor to many of these modern deep-learning powered NLP systems is the availability of pre-trained concept representations. These concept representations are learned such that the precise syntactic and semantic relationships between concepts are preserved in the learned vector space. Due to its wide usability and practical implications, learning high-quality concept representations remains a fundamental problem in the research area of biomedical NLP. Over the past few years, many biomedical word embedding models such as BioWordVec [124], BioBERT [2], and SciBERT [125] have been proposed in the literature. Recent trends show that the contextualized embedding approaches (i.e., BERT [2] based models) significantly outperforms the conventional word embedding approaches (i.e, Word2vec [16] based models) in a variety of NLP tasks. This is primarily because the contextual embedding approaches are able to capture the semantic properties of concepts under diverse linguistic contexts. Despite significant advances made, the contextual embedding approaches suffer from high computational and space costs. For instance, BioBERT takes nearly 23 days to train on the entire biomedical corpora and requires hundreds

of millions of parameters [2]. This is limiting for rapidly evolving domains such as biomedicine (around 3,000 articles are added every-day [2]) wherein the timely update of concept representations is essential to reflect the accurate knowledge of the field. Moreover, such longer training times severely impacts the practicalities of contextual embedding approaches in time-critical medical applications such as real-time disease diagnostics and monitoring [126]. To address these issues, it is imperative to develop representation learning approaches such that the contextual embedding models are able to efficiently (yet accurately) adapt the feature representations of biomedical concepts to the progressively available data. This is the crux of the problem that this paper attempts to address. Prior research has attempted to accelerate the efficiency of embedding models through a range of solutions such as knowledge distillation [127], weights pruning [128], and continual learning [116]. Amongst them, the continual learning (CL) based approaches have attracted increasing attention due to their natural ability to adapt the representations to the continuous streams of data. However, the existing CL approaches [129, 116] have been predominantly designed for the embedding models proposed in the research area of computer vision. Directly applying these approaches to the NLP focused embedding models yields unsatisfactory performance due to the fundamental differences in the characteristics of imaging and textual datasets. Moreover, the specialized nature of the biomedical domain presents unique opportunities/challenges to leverage the rich semantic knowledge present in curated knowledge-bases (KB's) whilst designing an efficient representation learning approach.

To address the aforementioned challenges, we leverage upon the principles of continual machine learning [116, 129] and propose a new representation learning approach that efficiently yet accurately adapts the concept representations to the newly available data. Specifically, the proposed approach considers the successive corpus snapshots as a sequence of related tasks and updates the concept representations that are affected by the new snapshot, while preserving those that are well-trained previously. The main challenge in this strategy is to automatically identify the concepts whose representations are subject to retraining or retention, referred to as 'selective retraining' in the continual-learning literature [116, 129]. To address this, we propose a knowledge-guided retraining scheme wherein at every new corpus snapshot, we leverage the semantic knowledge from KB's to identify and retrain the representations of those concepts whose corpus-specific context evolved coherently with-respect-to their KB-specific context. More concretely, using the curated context information from KB's

as a reliable signal, we discern the coherency/noisiness of the concept's corpus-specific contextual neighbors and retrain/retain their feature representations accordingly. Following this strategy, the proposed knowledge-guided technique is iteratively applied to the consecutive snapshots, and the concept representations are generated efficiently. Furthermore, we propose a knowledge-guided pruning mechanism that eliminates the redundant parameters present in the overparameterized transformer-based embedding architectures [130, 131], thus greatly improving their overall memory efficiency. Finally, the proposed approach is designed to remain agnostic to the choice of the embedding loss-function. Given the fact that there are multiple competing contextualized embedding models such as BioBERT [2], SciBERT [125], and ClinicalBERT [132] to generate the concept representations, it is desirable to develop approaches that do not jeopardize the embedding training process, and flexibly enables the users to utilize the proposed technique as a pluggable module for obtaining the improved training performance.

In this research, our contributions can be summarized as:

- We propose a new representation learning approach that efficiently (yet accurately) adapts the concept representations to the newly available data. While the methods proposed in this research are entirely general, our focus on the biomedical domain has immediate practical benefits for the practitioners of biomedical data science.

- The proposed research explores the usefulness of semantic knowledge present in the curated KB's from a new perspective, i.e, improving the training efficiency of embedding models. To achieve this, we conceptualize the paradigm of knowledge-guided continual learning and design new techniques such as knowledge-guided retraining and pruning.

- Extensive experiments in datasets from four bioNLP tasks demonstrate that the proposed approach can significantly improve the computational performance of the state-of-the-art biomedical word embedding models.

6.2 Related Work

6.2.1 Word Embedding In Biomedicine

Learning meaningful representations of concepts is a fundamental problem in the research area of biomedical NLP. For a recent survey on this topic, please refer [133]. Over

the past decade, many approaches [16, 65, 134, 135, 136, 17] for deriving word embeddings have been proposed in the literature. Among them, the prediction-based [16]) and count-based [17] models attracted significant attention from the research community. Prediction based approaches derive the word embeddings by optimizing the language model objectives that predict the next word given its context. In contrast, the count based models exploit the global word-context co-occurrence counts to obtain the word representations. Apart from capturing the implicit semantics, these approaches produce special analogical relations that are useful for various practical tasks. More recently, contextualized representation learning approaches (i.e., BERT based models) have obtained state-of-the-art performance in a number of bioNLP tasks such text classification, document retrieval and question-answering. Unlike conventional embedding approach such as word2vec, these approaches capture the semantic properties of concepts under varying local-contexts. Due to their promising results, a number of BERT based models such as BioBERT [2], SciBERT [125], ClinicalBERT [132], BlueBERT [137] and PubMedBERT [138] have been proposed. Despite the significant accuracy gains achieved by these embedding models, they still incur significant costs both in terms of training-time and memory. Thus, it is desirable to develop efficient representation learning approaches that can accelerate both their training-time and memory efficiency.

6.2.2 Continual Learning

Continual learning (CL) [116, 129] is a special type of online learning that incrementally acquires and fine-tunes information from non-stationary data distributions. The main challenge in CL is to continually accommodate the new information from streams of data while retaining the past knowledge. To address this, the CL based approaches perform selective retraining such that the network parameters adapt to the new information without overwriting the previously consolidated knowledge. In the recent years, various approaches such as regularization [129], dynamic architecture [116], and memory replay [139] has been proposed to tackle the issue of selective retraining. [129] proposed elastic weight consolidation (EWC) model that regularizes the model parameters at each step via fisher information matrix. [140] proposed to incrementally train an autoencoder by adding in new neurons for a group of difficult examples with high loss, and later merging them with other neurons to prevent redundancy. [116] proposed a

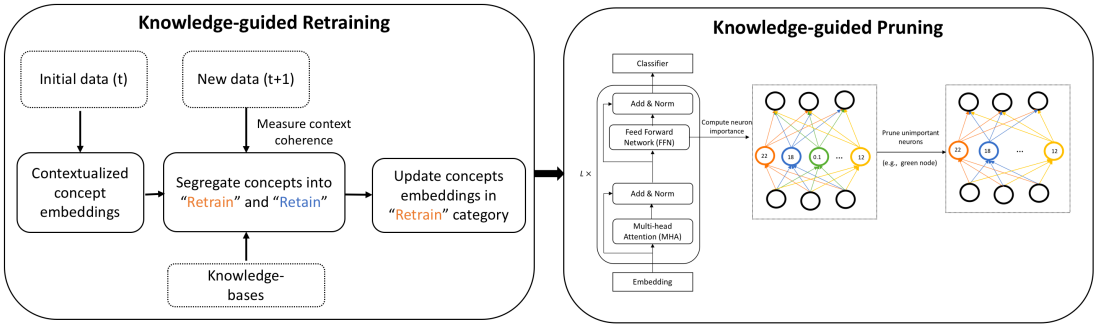


Figure 6.1: Overview of the proposed knowledge-guided retraining and pruning approach.

model that combines the best of both architectural and regularization strategies. Different from these approaches, the authors in [141] proposed to block any changes to the model trained on previous knowledge and expand the network architecture by allocating sub-networks with fixed capacity to be trained with the new information. While the aforementioned approaches made significant advances, none of them explored the usefulness of semantic knowledge present in curated KB's to design a knowledge-guided continual learning approach.

6.3 Methodology

We consider the problem of efficiently learning the representations of biomedical concepts under the continual learning scenario, where the corpus-snapshots (or training data) arrive at the model in a sequence. Specifically, our goal is to incrementally learn the representations from a sequence of T corpus-snapshots, $t = 1, \ldots, T$ for unbounded T. Each time-slice t comes with a training dataset $\mathcal{D}_t$, and all the previous datasets up to $t-1$ are not available. We consider an overall vocabulary $\mathcal{V} = \{w_1, \ldots, w_V\}$ of size $|V|$. $\mathbf{w} \in \mathbb{R}^d$ denotes a $d-$dimensional word embeddings that can be derived from any contextualized word embedding models (e.g., BioBERT [2], SciBERT [125] or ClinicalBERT [132]). Since the contextualized word embedding models generate a context-specific representations for each word, the representations for each word w specific to the context S is denoted as $E(\mathbf{w}, S)$. E refers to the chosen contextualized embedding model. Now, given a pre-trained contextual embeddings at initial time $t = 1$, i.e., $E(\mathbf{w}^t, S^t)$, the objective is to efficiently update the representations for each word w

at successive time-slices, i.e., $E(\mathbf{w}^{t+1}, S^{t+1}), \ldots, E(\mathbf{w}^{T}, S^{T})$. To achieve this, we propose two strategies: a) knowledge-guided retraining (Section 6.3.2), and b) knowledge-guided pruning (Section 6.3.3). These strategies efficiently adapt the concept representations to the newly available data whilst maintaining its prediction accuracy. Figure 7.2 shows an overview of the proposed knowledge-guided approach.

6.3.1 Preliminaries

Transformers

Transformer architecture [142] underscores the foundation behind BERT-based models. Basically, a transformer is a stack of layers composed of a multi-head attention and a feed-forward network. The multi-head attention layer consists of multiple attention heads that are executed in parallel. Each attention head takes matrix $\mathbf{X}$ where each row represents an element of the input sequence, and updates their representations by aggregating information from their context using the attention mechanism [143].

$$\mathbf{Z} = Softmax(\mathbf{X}^T\mathbf{Y}(\mathbf{QX} + \mathbf{P}))\mathbf{WX}, \tag{6.1}$$

where $\mathbf{Y}$, $\mathbf{W}$, $\mathbf{Q}$ and $\mathbf{P}$ are the matrices of parameters. The outputs from these heads are concatenated along the time-steps into a sequence of vectors. Then, a fully connected feedforward network is applied to each element of this sequence independently. Both of these layers are followed by an $AddNorm$ operation that consists of a residual connection and a layer normalization.

Unified Medical Language System Definitions

Unified Medical Language System (UMLS [144]) is an integrated biomedical knowledge resource that provides definitions associated with the medical concepts. As these definitions are curated by the subject-matter-experts, they are considered to be highly accurate. Table 6.1 presents an example of the UMLS definition for 'coronavirus infections'.

Medical concept hierarchy

Medical concepts in the biomedical domain are arranged in an hierarchical fashion [145] (i.e., *ISA* tree). The distance between concepts in the tree indicates the degree of semantic proximity between them. The depth of a concept in the tree indicates its level of specificity.

Table 6.1: Example of a medical concept and its definition obtained from the UMLS.

Medical Concept	Definition
Coronavirus Infections	Virus diseases caused by the coronavirus genus. Some specifics include transmissible enteritis of turkeys (enteritis, transmissible, of turkeys) and feline infectious peritonitis and transmissible gastroenteritis of swine (gastroenteritis, transmissible, of swine)

6.3.2 Knowledge-guided Retraining

Our main objective is to identify the concepts that require the retraining of their feature representations in order to accommodate the newly available data. To accomplish this, we quantify the context-coherence of concepts (between their current corpus-specific context and KB-specific context), and update the representations of those concepts whose context evolved coherently over time. The rationale is the following: since the expert-curated context information available from KB's are both accurate and stable, they provide reliable feedback to discern whether the current corpus-specific context information is coherent or noisy. If coherent, we retrain the representations of concepts. Otherwise, we retain the previously well-trained representations. This context-coherence based premise is supported by recent research [146] that have shown that the quality of contextual neighbors significantly contributes to the stability/quality of medical concept embeddings. To this end, we propose to measure the context-coherence of concepts from two perspectives: a) explicitly shared context between corpus and KB, and b) implicitly shared context between corpus and KB.

Explicitly Shared Context.

Let $S^t_{cp}(w)$ and $S^t_{umls}(w)$ denote the set of corpus-based and UMLS-based (refer Section 6.3.1) context terms for a concept w at time t respectively. To measure the coherence between these context sets, we first compute the intersection $S^t_{cp}(w) \cap S^t_{umls}(w)$ of their shared concepts. However, this straightforward mechanism severely penalizes the inexact concept-name matches across the two sets. To mitigate this issue, we augment the context-sets with the semantic neighbours of concepts from the medical concept

hierarchy (refer Section 6.3.1). Then, we use the cartesian product of the two context sets $S^t_{cp}(w) \times S^t_{umls}(w)$ to determine the pairs of concepts that adequately indicate relatedness between the context-sets. Pairs of concepts (m^t_{cp}, m^t_{umls}), where $m^t_{cp} \in S^t_{cp}(w)$ and $m^t_{umls} \in S^t_{umls}(w)$, whose similarity is above pre-defined threshold of semantic similarity are retained, and those below are discarded. To compute the semantic similarity between concepts the measure of dice similarity is used. Dice similarity computes the proportion of common ancestors between the concepts in the medical hierarchy, thus accounting for their shared semantics at a granular level. For two concepts m^t_{cp} and m^t_{umls} the dice similarity is computed as:

$$dice(m^t_{cp}, m^t_{umls}) = 2 \times \frac{|ancestors(m^t_{cp}) \cap ancestors(m^t_{umls})|}{|ancestors(m^t_{cp})| + |ancestors(m^t_{umls})|} \tag{6.2}$$

where $ancestors(m^t_{cp})$, $ancestors(m^t_{umls})$ refers to the set of all ancestors of m^t_{cp} and m^t_{umls} in the medical concept hierarchy respectively. The maximum similarity between two concepts computed using dice similarity is 1 (i.e., $m^t_{cp} = m^t_{umls}$). The range of similarity values is $[0, 1]$. Pairs of concepts whose dice similarity exceed the threshold of semantic similarity (manually assigned as $\tau_{sim} = 0.75$) are normalized to a value of 1. The normalized dice similarity is:

$$dice_C(m^t_{cp}, m^t_{umls}) = \begin{cases} 1 & \text{if } dice(m^t_{cp}, m^t_{umls}) > \tau_{sim} \\ 0 & \text{otherwise} \end{cases} \tag{6.3}$$

where C denotes the total number of concept pairs. The semantic relatedness (sr') between $S^t_{cp}(w)$ and $S^t_{umls}(w)$ is the sum of the normalized pairwise dice similarity scores that exceed the threshold of semantic similarity across the cartesian product of the context sets $S^t_{cp}(w) \times S^t_{umls}(w)$.

$$sr'(S^t_{cp}(w), S^t_{umls}(w)) = \sum_{(a,b) \in S^t_{cp}(w) \times S^t_{umls}(w)} dice_C(a, b) \tag{6.4}$$

An outcome of this semantics-enhanced shared context metric is that a broad range of semantic relatedness scores between context-sets may exist. To dampen the major differences in similarity scores of different context sets, we apply a log reduction on the normalized dice similarity scores. This is achieved by first computing the relatedness score between a given concept in context set $S^t_{cp}(w)$ against the entire set of concepts in the context set $S^t_{umls}(w)$. This calculation yields the similarity score:

$$sim'(a, S^t_{umls}(w)) = \sum_{b \in S^t_{umls}(w)} dice_C(a, b) \tag{6.5}$$

The log reduction is then applied to $sim'(a, S^t_{umls}(w))$, and the overall semantic relatedness between the two context sets is the aggregate of the log-reduced scores for

each concept in $S^t_{cp}(w)$ and the entire set in $S^t_{umls}(w)$. Below is the resultant metric to calculate the explicitly shared context:

$$sim''(S^t_{cp}(w), S^t_{umls}(w)) = \sum_{a \in S^t_{cp}(w)} log(1 + sim'(a, S^t_{umls}(w))) \tag{6.6}$$

Implicitly Shared Context.

While the technique proposed in Section 6.3.2 captures the explicitly shared context, it misses to factor in the implicit semantics. Thus, from a complementary perspective, we propose to quantify the context-coherence by measuring the amount of implicit semantics shared between the concepts corpus and KB-specific context-sets. To accomplish this, we first use the BERT [142] language model to produce the representations of concepts from their UMLs definitions alone. Then, we cluster the representations of each unique concept into a number of partitions. Note that since BERT produces context-specific representations, a concept can have multiple representations. The cluster partitions capture the usages of the concepts along different dimensions (or senses). For example, the three clusters for the medical concept 'coronavirus infection' are 'disease', 'virus', and 'infections'. These dimensions are formally referred to as the semantic types [144] in the biomedical domain. The semantic types basically provide a formal categorization of the biomedical concepts. Next, similar to the clustering mechanism of KB-specific representations, we cluster the initial corpus-specific representations ($t = 1$) as well. Now, given the clustered usage representations from both corpus and KB, we use the Jensen-Shannon divergence (JSD) [147] to compute the implicit context-coherence of concepts at any time t. Specifically, we count the number of occurrences of each usage type l in a given time slice t (we refer to this count as $freq(l, t)$) and obtain frequency distributions $\mathbf{f}^t_w$ for each time-slice as the following:

$$\mathbf{f}^t_w[l] = freq(l, t) \quad l \in [1, L_w] \tag{6.7}$$

where L_w denotes the total number of usage types for concept w. This normalised frequency distributions can be interpreted as probability distributions over usage types $\mathbf{u}^t_w : \mathbf{u}^t_w[l] = \frac{1}{V_t}\mathbf{f}^t_w[l]$. Below is the metric used to compute the implicit coherence:

$$JSD(\mathbf{u}^{cp}_w, \mathbf{u}^{kb}_w) = H\left(\frac{1}{2}\left(\mathbf{u}^{cp}_w + \mathbf{u}^{kb}_w\right)\right) - \frac{1}{2}\left(H\left(\mathbf{u}^{cp}_w\right) - H\left(\mathbf{u}^{kb}_w\right)\right) \tag{6.8}$$

where H is the normalized shannon entropy [148]. By quantifying the implicitly shared context, the proposed approach is able to measure the context coherence of concepts at a granular level.

Joint Context-coherence.

Both the explicitly and implicitly shared-context measures the context-coherence of concepts between corpus and KB from complementary perspectives. Thus, we compute the coherence score of concepts in a joint manner.

$$score(w) = (1 - \alpha) \cdot \ explicitcontext(w) + \alpha \cdot implicitcontext(w) \tag{6.9}$$

where $explicitcontext(w), implicitcontext(w)$ are computed using Equations 6.6 and 6.8 respectively. The value of α controls the contribution of each part. Using this measure, the concepts with high values of coherence score are chosen for retraining their representations in the consecutive snapshots, while those with lower coherence values (i.e., noisy context) are simply retained. In this way, the proposed knowledge-guided selective retraining strategy is able to produce the representations promptly, thereby significantly accelerating the training time of contextualized embedding models.

6.3.3 Knowledge-guided Pruning

While the proposed knowledge-guided retraining scheme accelerates the training time, it is equally important to address the memory challenges posed by the contextualized word embedding models. Specifically, the contextualized embedding models are recognized to be overparameterized [131, 130], which makes them memory inefficient for applications requiring execution in real-time. To address this issue, we propose a knowledge-guided pruning mechanism that localizes the knowledge (or important parameters) within the network layers and prunes the redundant parameters such that only the relevant knowledge is retained. Concretely, the proposed approach explores the recent developments in block structured pruning techniques [128], and designs a new objective that effectively eliminates the uncritical weights (in groups) in transformer-based contextualized embedding models whilst preventing the possible information loss. Figure 7.2 (right side) shows an overview of the proposed knowledge-guided pruning approach.

Since almost all the state-of-art contextualized embedding models [2, 125, 132] employ the transformer architecture [142] at its core, we use them as our running example for the proposed pruning scheme. Consider an N-layer transformer where the weights of the n-th layer are denoted as $\boldsymbol{\theta}_n$. The function $f\left(\left\{\boldsymbol{\theta}_n\right\}_{n=1}^{N}, \mathcal{D}_t\right)$ denotes the loss function of the chosen contextualized embedding model. $\mathcal{D}_t$ denotes the training data at time t. To efficiently localize the knowledge in network layers, the weight matrix $\boldsymbol{\theta}_n$ is divided into K blocks $\boldsymbol{\theta}_n = [\boldsymbol{\theta}_{n1}, \boldsymbol{\theta}_{n2}, \cdot, \boldsymbol{\theta}_{nK}]$, where $\boldsymbol{\theta}_{nm} \in \mathbb{R}^{n \times m}$. Let $[\boldsymbol{\theta}_{nm}]p, :$ and $[\boldsymbol{\theta}_{nm}]\ :, q$

denote the p-th row and the q-th column of $\boldsymbol{\theta}_{nm}$ respectively. For each row/column block, we compute the parameters importance with respect to the loss function and prune the weights that are below a preset threshold. Specifically, our objective is to reduce the number of columns and rows in the blocks of weight matrix whilst maintaining the prediction performance.

$$\begin{aligned} &\textit{minimize} \quad f\big(\{\boldsymbol{\theta}_n\}_{n=1}^{N}, \mathcal{D}_t\big) \\ &\textit{subject to} \quad \#\text{ of non-zero block rows in } \boldsymbol{\theta}_n \text{ is less than } r_n \\ &\textit{subject to} \quad \#\text{ of non-zero block rows in } \boldsymbol{\theta}_n \text{ is less than } c_n \end{aligned} \tag{6.10}$$

where r_n and c_n are the desired non-zero block rows and columns respectively. Moreover, to effectively prune the weights in groups, we add the group lasso regularization [149] to the objective function. The objective function becomes:

$$min \quad f\big(\{\boldsymbol{\theta}_n\}_{n=1}^{N}, \mathcal{D}_t\big) + \lambda_1 \sum_{i=1}^{N}\sum_{j=1}^{K} ||\boldsymbol{\theta}_{ij}||_g \tag{6.11}$$

where λ_1 controls the relative strength of lasso regularization, and $||\cdot||_g$ denotes group lasso regularization. The groups are defined based on the incoming weights for each neuron in the feed forward network layer of the transformer architecture. While the group lasso regularization significantly promotes the structured sparsity, its precise application is needed to address the overparameterized nature of transformer-based embedding architectures. Thus, we introduce a new loss that penalizes the weights that are neither close to 0 nor 1, pushing them close to either 0 or 1. The introduced loss is shown below:

$$\sum_{i=1}^{N}\sum_{j=1}^{K} \big(\boldsymbol{\theta}_{ij} \times (1 - \boldsymbol{\theta}_{ij})\big) \tag{6.12}$$

The key addition of this loss to the objective function facilitates the precise application of group lasso such that it effectively zeros out the non-critical weights. Next, we add this loss to the block-based row pruning and column pruning formulation. For block-based row pruning, we solve:

$$min\ f\big(\{\boldsymbol{\theta}_n\}_{n=1}^{N}, \mathcal{D}_t\big) + \lambda_1 \sum_{i=1}^{N}\sum_{j=1}^{K} |||[\boldsymbol{\theta}_{ij}]_{p,:}||_2 + \lambda_2 \sum_{i=1}^{N}\sum_{j=1}^{K} \big(\boldsymbol{\theta}_{ij} \times (1 - \boldsymbol{\theta}_{ij})\big) \tag{6.13}$$

For block-based column pruning, we solve:

$$min\ f\big(\{\boldsymbol{\theta}_n\}_{n=1}^{N}, \mathcal{D}_t\big) + \lambda_1 \sum_{i=1}^{N}\sum_{j=1}^{K} |||[\boldsymbol{\theta}_{ij}]_{:,q}||_2 + \lambda_2 \sum_{i=1}^{N}\sum_{j=1}^{K} \big(\boldsymbol{\theta}_{ij} \times (1 - \boldsymbol{\theta}_{ij})\big) \tag{6.14}$$

In Equations 6.13 and 6.14, λ_1 and λ_2 control the contribution of each part. Using this strategy, the proposed knowledge-guided pruning mechanism is able to significantly

Algorithm 1 KNOWLEDGE-GUIDED PRUNING

Input: Pretrained transformer model with weight matrix $\boldsymbol{\theta}$, threshold ϵ
Output: Pruned weight matrix $\boldsymbol{\theta_s}$
Initialize $\boldsymbol{\theta_s} = \boldsymbol{\theta}$
Divide $\boldsymbol{\theta}_s$ into K matrices: $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, .., \boldsymbol{\theta}_K$
Set $i = 1$
Set total number of iterations = $maxT$
Solve the regularization problem (13),(14) using ADAM
while i $\leq$ maxT **do**
    $l_2\_norms_l$ equals the l_2 norm of each p-th/q-th row of $\boldsymbol{\theta}_s$
    if $l_2\_norms_l \leq \epsilon$ **then**
        $\boldsymbol{\theta}_s(p, :) = 0$
        $\boldsymbol{\theta}_s(:, q) = 0$
    end if
end while
$\boldsymbol{\theta}_s = concatenate\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2..., \boldsymbol{\theta}_K\}$

improve the memory efficiency of transformer-based embedding models, and at the same time maintain its predictive accuracy. The pseudocode of the proposed approach is shown in Algorithm 3.

6.4 Experiments

Our main objective is to examine the capability of proposed approach to accelerate the learning efficiency of contextualized embedding models whilst maintaining their prediction accuracy. As such, we compare the quality of continual (or incremental) embeddings produced by our approach for contextualized embedding models [2, 125, 132, 138] with their batch-mode counterparts in terms of both accuracy and computational efficiency.

6.4.1 Datasets

Pre-Training Data. We use the abstracts from PubMed [3] (1960-2019) as our training dataset to generate the concept representations. PubMed contains more than 30 million articles from the areas of life-sciences and biomedicine. We follow the preprocessing steps

suggested in BioBERT [2] to generate an overall vocabulary of over 3.2 billion concepts. To generate the incremental embeddings, the preprocessed dataset is split yearly (i.e., publication year) and the proposed approach is applied. The batch-mode embeddings are generated from the same dataset, i.e., PubMed (1960-2019), using the source code of the models from their public releases. For a head-to-head comparison, we set the hyperparameter values of the models as reported in their respective papers.

BioNLP Tasks. The experiments are conducted on four biomedical NLP tasks that have publicly available datasets, including named entity recognition, relation extraction, sentence similarity, and question answering. Below are detailed descriptions for each task and their corresponding datasets.

Named Entity Recognition. For the biomedical named entity recognition task, we conduct experiments on the Natural Center for Biotechnology Information Disease (NCBI) [150] and Biocreative II Gene Mention (BC2GM) datasets [151]. NCBI dataset contains 793 abstracts with 6892 annotated disease mentions. BC2GM consists of sentences with manually labeled gene and alternative gene entities. We use the pre-processed version of train, development, test splits released by [152].

Relation Extraction. For this task, we consider the drug-drug interaction (DDI) [153] and chemical protein interaction (ChemProt) datasets released by [154]. These datasets contain sentence-level annotation of drug-drug interactions and protein–chemical relations respectively. We follow the pre-processing procedure described in [155] to reduce noise in the dataset.

Sentence Similarity. BIOSSES [156] is a sentence similarity dataset that consists of 100 pairs of sentences. These sentences are annotated by five subject-matter-experts with a similarity score in the range of 0 (no relation) to 4 (equivalent meanings). Another dataset, MedSTS [157] consists of 1,068 sentence pairs that are annotated by two experts. The similarity of sentence pairs is annotated in terms of five categories from 0 (not similar) to 5 (very similar). The average score from the annotators is considered as the final score.

Question Answering (QA). For the task of QA, we use the BioASQ factoid dataset [158]. Each question is paired with a reference text that contains multiple sentences and a yes/no answer. Similarly, another chosen QA dataset (PubMedQA [159]) contains a set of questions, each with a reference text, and an annotated label of whether the text contains the answer to the question (yes/maybe/no).

6.4.2 Experimental Setup

To fine-tune the models on downstream tasks, we add a single linear layer (or regression layer for sentence similarity) on top of each contextualized embedding models. We train the embedding models on bioNLP tasks with their corresponding datasets. This training procedure flexibly adapts the embedding models to specific tasks. Following the practice in contextual embedding literature [2, 125, 132], the range of hyperparameters are chosen to be the following: learning rate within the range [1e-5, 3e-5, 5e-5], batch size [10, 16, 32, 64] and epoch number [2–60]. Considering the average prediction performance, the learning rate, batch size, and epochs are set to 3e-5, 32 and 4 respectively on all four tasks. Analyzing the contributions of explicit and implicit context (details in Section 6.4.4), the value of α in Equation 6.9 is set to 0.4 and 0.6 respectively. Similarly, the values of both λ_1 and λ_2 in Equations 6.13 and 6.14 are set to 0.5.

Baseline Models. Our baseline models are the batch versions of BioBERT [2], SciBERT [125], ClinicalBERT [132], BlueBERT [137] and PubMedBERT [138]. As shown in Table 6.2, for each contextualized embedding model, we report the accuracy and efficiency metrics for both the original (batch) and proposed (continual) versions. The continual version of the embedding models are named with suffix "-CL" in the Tables 6.2 and 6.3. Moreover, we compare our results with the existing efficient transformer-based models such as [127, 160], and continual learning approaches such as [129, 116, 161].

Evaluation Metrics. To measure the quality of concept embeddings in downstream bioNLP tasks, we follow the convention in literature [2, 125, 132] and report micro F1 for named-entity-recognition and relation extraction, Pearson coefficient for sentence similarity, and accuracy for question-answering. For measuring the computational efficiency, we report floating-point operations (FLOPs [127]). Specifically, FLOPs calculate the number of floating-point operations that the models perform for a single process.

6.4.3 Results

Tables 6.2, 6.3, 6.4, and 6.5 report the results of the proposed approach and baseline algorithms in the datasets from four bioNLP tasks. In Table 6.2, we compare the performance of the contextualized embedding models that are obtained after applying the proposed selective retraining technique with their original batch counterparts. Our approach performs on par with the batch versions in terms of prediction accuracy whilst

Table 6.2: Comparison of prediction performance and training efficiency in the bioNLP datasets. The evaluation metric for NCBI, BC2GM, DDI, and ChemProt is micro-F1. BIOSSES and MedSTS use Pearson Coefficient. BioASQ and PubMedQA use Accuracy.

Model	NCBI	BC2GM	DDI	ChemProt	BIOSSES	MedSTS	BioASQ	PubMedQA	#Params	#FLOPs
BioBERT (Batch)	89.772	83.120	80.188	76.645	88.321	81.524	83.242	61.301	109M	22.5B
BioBERT-CL (**Proposed**)	89.778	83.123	80.189	76.646	88.323	81.525	83.246	61.312	**66.1M**	**13.7B**
SciBERT (Batch)	88.575	83.762	81.361	74.463	85.154	81.194	77.832	58.194	106M	19.2B
SciBERT-CL (**Proposed**)	88.579	83.765	81.365	74.467	85.155	81.196	77.838	58.199	**54.4M**	**11.4B**
ClinicalBERT (Batch)	86.242	80.562	77.898	72.542	90.182	79.301	67.322	48.186	103M	17.8B
ClinicalBERT-CL (**Proposed**)	86.245	80.566	77.899	72.544	90.185	79.307	67.326	48.188	**52.3M**	**9.4B**
BlueBERT (Batch)	88.143	81.172	76.686	70.712	84.495	76.145	69.788	47.572	107M	20.1B
BlueBERT-CL (**Proposed**)	88.145	81.177	76.689	70.717	84.497	76.146	69.784	47.575	**55.6M**	**12.5B**
PubMedBERT (Batch)	87.383	85.122	81.761	76.454	91.205	82.643	86.766	54.192	92M	14.2B
PubMedBERT-CL (**Proposed**)	87.388	85.127	81.766	76.457	91.208	82.642	86.767	54.198	**40.3M**	**5.3B**

Table 6.3: Comparison of prediction performance and compression rate (memory) in the bioNLP datasets.

Model	NCBI	BC2GM	DDI	ChemProt	BIOSSES	MedSTS	BioASQ	PubMedQA	Rate
BioBERT (Batch)	89.772	83.120	80.188	76.645	88.321	81.524	83.242	61.301	N/A
BioBERT-CL (Proposed)	90.971	85.220	82.294	77.545	89.684	83.951	84.321	63.689	1.487×
SciBERT (Batch)	88.575	83.762	81.361	74.463	85.154	81.194	77.832	58.194	N/A
SciBERT-CL (Proposed)	89.244	84.271	81.698	75.690	85.435	83.247	78.856	58.411	1.487×
ClinicalBERT (Batch)	86.245	80.566	77.899	72.544	90.185	79.307	67.326	48.188	N/A
ClinicalBERT-CL (Proposed)	87.321	81.288	78.157	72.955	91.974	79.922	68.431	49.242	1.487×
BlueBERT (Batch)	88.145	81.177	76.689	70.717	84.497	76.145	69.788	47.575	N/A
BlueBERT-CL (Proposed)	89.257	82.342	77.667	71.832	85.525	76.796	69.924	48.691	1.831×
PubMedBERT (Batch)	87.383	85.122	81.761	76.454	91.205	82.643	86.766	54.192	N/A
PubMedBERT-CL (Proposed)	88.145	86.238	82.789	77.322	92.912	84.516	87.142	56.289	1.667×
DistilBERT (Batch)	85.111	83.122	80.761	74.454	86.205	78.643	82.766	51.192	N/A
DistilBERT-CL (Proposed)	85.997	84.382	81.008	75.831	87.102	80.011	83.381	52.111	1.667×
FastBERT (Batch)	83.001	81.222	78.112	72.121	88.102	78.129	83.112	53.225	N/A
FastBERT-CL (Proposed)	84.481	82.092	79.119	73.311	89.999	79.587	84.322	54.981	1.831×

Table 6.4: Comparing pruning results of BioBERT with different compression rates.

Compression Rate	NCBI	BC2GM	DDI	ChemProt	BIOSSES	MedSTS	BioASQ	PubMedQA
1×	89.998	83.345	81.378	76.811	89.859	81.117	84.113	62.229
1.48×	89.772	83.120	80.188	76.645	88.321	81.524	83.242	61.301
2.0×	87.303	82.183	77.472	75.128	88.225	78.384	82.295	60.689
4.0×	85.903	79.133	74.001	71.204	88.102	73.193	81.295	59.293

Table 6.5: Comparing prediction performance with different continual learning methods in the bioNLP datasets.

Model	NCBI	BC2GM	DDI	ChemProt	BIOSSES	MedSTS	BioASQ	PubMedQA
Elastic Weight Consolidation [129]	81.146	72.156	69.056	67.802	71.492	65.034	68.175	45.827
Dynamically Expandable Networks [116]	84.289	76.122	73.095	69.679	74.671	69.788	71.992	49.881
AdapterBERT [161]	86.557	79.888	76.231	71.708	84.401	75.809	76.101	55.223
Proposed	**89.778**	**83.123**	**80.189**	**76.646**	**88.323**	**81.525**	**83.246**	**61.312**

significantly improving their computational efficiency. This result validates the effectiveness of the proposed selective retraining scheme that retrains the representations of only those concepts whose context evolved coherently with respect to their KB-specific curated context. Table 6.3 reports the performance of the proposed approach after applying the knowledge-guided pruning mechanism. We set a compression rate of 1.428× (i.e., 30% sparsity) or above for all the models. The results show that the proposed technique is capable of improving memory efficiency while maintaining the prediction accuracy. Notably, on a majority of datasets, the pruned models further improve the overall accuracy. We attribute this result to the effective pruning of redundant weights in the transformer-based embedding architecture. We also compare the performance with compact language models such as DistilBERT [127] and FastBERT [160]. The proposed knowledge-guided techniques are able to boost their computational efficiency. Analyzing the results further, we evaluate the performance changes (using BioBERT) with varying compression rates and report the results in Table 6.4. Results show that the performance varies significantly under different levels of compression rates. In general, as we increase the compression rate beyond a certain threshold, the performance starts to degrade. For specific tasks such as BIOSSES, we can achieve up to 4 compression rates from the baseline model with almost zero performance loss. Results on tasks such as BIOASQ and PubMedQA show minor degradation, while the results on NCBI, BC2GM, DDI, ChemProt, and MedSTS show higher degradation when the compression rate is set to 4.0. We speculate the different performance results due to the differences in the characteristics of datasets and the unique challenges posed by the particular tasks. Next, to analyze the effectiveness of the proposed continual learning approach, we compare it with three existing methods that are applicable to our problem setting. Note that the continual learning techniques are mainly designed for multi-task learning problems, and thus a direct comparison with a majority of techniques cannot

be performed. While our research builds upon the ideas of continual learning, it is designed for single task incremental settings. Table 6.5 reports the results. The proposed knowledge-guided continual learning approach outperforms the traditional CL approaches. We attribute two reasons for this: a) The existing computer vision focused approaches do not effectively characterize the syntax and semantics information present in natural language text, b) The existing approach are inept at exploiting the semantic knowledge present in curated KB's.

6.4.4 Ablation Studies

In this section, we perform ablation studies over several parameters to better understand their relative influence. The parameters chosen are the following: the effect of explicit and implicit context in the knowledge-guided retraining scheme, effect of regularization losses in the objective function of Equation 6.13, and the numbers of blocks in the knowledge-guided pruning phase. We chose the datasets from each of the four tasks and report the results using BioBERT [2]. The evaluation metrics are F1-score for NCBI and DDI, Pearson correlation for BIOSSES, and accuracy for PubMedQA.

Effect of explicit and implicit context in context-coherence

To understand the relative influence of explicit and implicit context in the overall context-coherence score, we perform a component-wise analysis. Table 6.6 presents the results. As it can be observed, the prediction performance is best when both the components are exploited jointly. Notably, the contribution of implicit context is greater than that of the explicit context. We believe that this is due to the capability of the proposed implicit metric to capture the underlying semantics of biomedical text data at a granular level.

Effect of regularization parts in knowledge-guided pruning

The regularization parts in the objective function (Equation 6.13) of knowledge-guided pruning are added to eliminate the unimportant parameters. Table 6.7 summarizes the results of both the losses. While the group lasso regularization obtains reasonable performance in the datasets, the addition of proposed loss (i.e., Equation 6.12) facilitates its application to be more precise.

Table 6.6: Influence of explicit and implicit context on the datasets from each of the four bioNLP task

Type of context	NCBI	DDI	BIOSSES	PubMedQA
Explicit context	82.128	76.284	84.194	55.486
Implicit context	84.485	78.382	86.103	58.983
Explicit+Implicit	89.778	80.189	88.323	61.312

Table 6.7: Influence of regularization parts on the datasets from each of the four bioNLP task

Regularization	NCBI	DDI	BIOSSES	PubMedQA
Group lasso	87.793	75.172	86.095	57.102
Proposed loss	86.834	76.933	86.119	58.502
Lasso+Proposed	90.971	82.294	89.684	63.689

Table 6.8: Influence of number of blocks in the proposed knowledge-guided pruning strategy

Number of blocks	8	128	256	768
BioBERT	85.329	86.392	88.756	90.971
SciBERT	82.378	83.481	85.912	89.244
ClinicalBERT	83.291	84.129	85.294	87.321
BlueBERT	85.183	86.299	87.447	88.143
PubMedBERT	84.692	85.391	86.566	87.388

Effect of number of blocks in knowledge-guided pruning

Table 6.8 presents the results of the prediction performance vs the number of blocks. The performance significantly increases with the number of blocks. This result indicates that the structured pruning based approaches have higher flexibility in exploring the sparsity of transformer-based embedding models.

6.5 Conclusion

In this research, we proposed a new representation learning approach that continually adapts the representations to the progressively available data. Specifically, the approach explores the semantic knowledge present in curated KB's to design a knowledge-guided strategy that selectively retrains the representations of those concepts whose context evolved coherently over time. Moreover, the proposed knowledge-guided pruning technique eliminates redundant parameters in the transformer-based models, thereby significantly improving its memory efficiency. Comprehensive experiments conducted in the datasets from four bioNLP tasks validate the efficacy of the proposed approach and demonstrates its potential usefulness in a variety of real-time biomedical applications.

Chapter 7

Continual Knowledge Infusion Into Biomedical Models

7.1 Introduction

Mining and analyzing the vast numbers of unstructured text in the biomedical domain offers great opportunities to advance scientific discovery [162]. Consequently, there is an increasing interest towards developing robust text-mining and natural language processing systems that can generate actionable insights and drive research frontiers. Many of these modern deep-learning powered bioNLP/text-mining systems utilize the pretrained feature representations of concepts as their input source. As such, numerous biomedical language models have been proposed in the machine learning literature. More recently, contextualized language models [2, 1] that capture the semantic properties of concepts under diverse linguistic contexts have achieved cutting-edge performance. Despite significant accuracy gains, these models are still unable to learn high-quality feature representations for concepts with low co-occurrence frequency (i.e., rare or domain-specific concepts). Such domain-specific concepts are abundantly present in the biomedical corpus and learning accurate representations for these concepts is essential to the success of predictive biomedical applications [163, 164, 136]. One way to address this challenge is by exploiting the complementary resources such as domain-expert curated knowledge-bases (KBs). Infusing the semantic knowledge from such KBs into the pretrained language models is likely to improve the representations for domain-specific concepts, and possibly even for those concepts that have adequate co-occurrence information. This is

the core objective of the proposed research in this paper.

Over the past few years, some efforts [165, 166, 167] have been dedicated to infusing the semantic knowledge into the contextualized language models. Despite their effectiveness, these approaches still have certain limitations. First, the existing approaches mainly incorporate the KBs by augmenting the language modeling objective with knowledge-specific regularizers. However, this strategy usually requires retraining of the entire model parameters that incurs significant computational overhead to the already overparameterized [130] contextualized embedding models. Second, the existing approaches are mainly designed for the general domain use-cases that primarily focused on integrating only one kind of KB, e.g., WordNet [168]. However, specialized domains such as biomedicine contain a plethora of well-organized KBs such as the Medical Subject Headings (MeSH) [145], International Classification of Diseases (ICD-10) [169], and Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) [170]. To flexibly incorporate multiple KBs, we propose to formulate the problem in a continual learning (CL) setting wherein the approach progressively integrates diverse semantic knowledge. Moreover, as the proposed CL formulation facilitates incremental updates of concept representations, it effectively mitigates the expensive retraining of contextualized language models. One critical issue in CL based formulation is to prevent catastrophic forgetting, i.e., the model abruptly forgets knowledge learned from previous KBs when learning on the new KB. To overcome this, we propose a new regularization mechanism that constraints the learned concept representations in the embedding space. This approach is different from the existing CL approaches [129, 171] that usually operate over the parameter space rather than the embedding space.

Meanwhile, as the majority of KBs in the biomedical domain are expressed as parent-child hierarchies, we focus on modeling them in this study. While some recently proposed language models [167, 166] have attempted to incorporate the hierarchical KBs, they still have certain drawbacks. Specifically, the existing approaches incorporate the hierarchical structure by solely modeling the direct hyponym-hypernym (i.e., one-level structure) relationships. This is limiting because these approaches miss to model the semantic contribution from concept's ancestors (i.e., multi-level structure). For instance, consider the concept "Heart Failure, Diastolic" shown in Figure 7.1. Existing approaches such as LIBERT [165] model the semantics of this concept by only considering its direct parent (i.e., Heart Failure). However, as it can be observed, the ancestors of this concept (i.e. "Heart Diseases", "Cardiovascular Diseases") provide useful semantic information too.

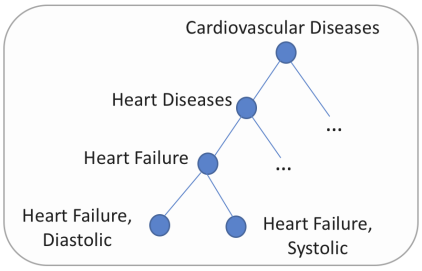


Figure 7.1: Example of a hierarchical structure extracted from MeSH taxonomy.

Thus, it is imperative to model the semantic contribution of ancestors in order to fully explore the comprehensive structure of hierarchical KBs. We propose to capture the semantic contribution of ancestors on a concept's representation via attention mechanism [143]. Furthermore, the existing approaches have largely missed to model the *distinctive* semantic information between siblings (i.e., concepts at the same level). For instance, although the concepts "Heart Failure, Diastolic" and "Heart Failure, Systolic" share the same parent, they add specific semantics to form their unique meaning. The proposed knowledge modelling strategy factors in this semantic information too.

Altogether, the proposed approach models the special topological properties of taxonomic KBs at a granular level, and develops a new continual learning based mechanism to integrate diverse KBs in a systematic manner. Finally, as the proposed approach does not change the core architecture of transformer based contextualized embedding models, it can be flexibly integrated with multiple competing pretrained biomedical language models such as BioBERT [2], SciBERT [125], and BioELMo [1] for boosting their overall prediction accuracy.

In this research, our contributions can be summarized as:

- The proposed approach integrates diverse KBs into the pretrained language models from a new perspective, i.e., in a continual fashion. Notably, the designed technique has special usability for the text-mining/NLP practitioners of the biomedical domain where a large number of KBs are known to exist.

- We propose a new technique to model the hierarchical structure of taxonomic KBs. By modeling the unique semantic contributions from both the ancestors and siblings, the proposed approach explores the taxonomic structure of KBs in a comprehensive manner.

- Extensive experiments in datasets from three bioNLP tasks demonstrate that the

knowledge-powered embedding can significantly improve the accuracy of state-of-the-art biomedical language models.

7.2 Related Work

7.2.1 Biomedical Language Models

Biomedical language models such as BioBERT [2], SciBERT [125], PubMedBERT [138], and BioELMo [1] have achieved cutting-edge performance in a variety of bioNLP tasks such as named-entity-recognition, relationship extraction, and question-answering. For a recent survey on this topic, please refer [172]. The initial language modelling approaches [51, 50] generated concept representations using models such as Skip-gram [18] and GLoVE [17]. Skip-gram based models learn concept representations by maximizing the probability of individual words given its context, whereas GLoVE based models minimize the reconstruction error between co-occurrence statistics predicted by the model and the global co-occurrence statistics observed in the training corpus. While these approaches were effective in practice, they generated context-agnostic representations, i.e., a single representation for each concept. Such decontextualized representations ignored the polysemous properties of words. To overcome this, recently proposed contextualized language models [2, 1] encode the semantics of concepts under varied linguistic contexts and generate context-specific representations. These context-sensitive approaches have demonstrated significant improvement in performance. Building upon the success of these approaches, some studies [173, 167, 165] have attempted to incorporate the prior knowledge into the pretrained language models. For instance, KnowBERT [167] incorporated synset-synset and lemma-lemma relationships from WordNet [168] into BERT. Similarly, LIBERT [165] injects hyponym-hypernym pairs present in the WordNet [168] into BERT. Despite important advances made by these approaches, none of them attempted to integrate multiple kinds of KBs in a continual manner.

7.2.2 Continual Machine Learning

Continual machine learning [129, 174] aims to train the models over a sequence of tasks in an online manner. It is often tackled as an online multi-task learning problem where the objective is to progressively accommodate the new knowledge while retaining the previously acquired knowledge. The main challenge in this learning paradigm is referred

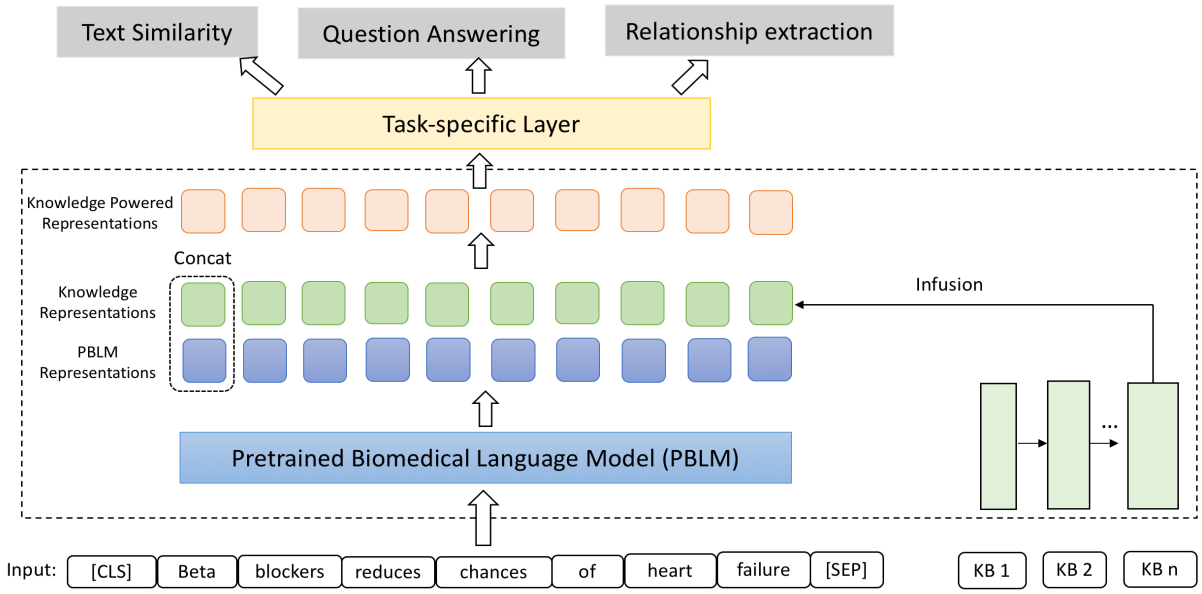


Figure 7.2: Continual Knowledge Infusion into the Pretrained Biomedical Language Models. PBLM refers to any pre-trained language model such as BioBERT [2].

to as catastrophic forgetting [129, 116], i.e., knowledge of previous tasks is abruptly forgotten when learning on the new task. Existing research has attempted to tackle this issue from three perspectives. The first class of approaches, i.e., memory-based approaches [174, 171] attempted to mitigate catastrophic forgetting by replaying the old training data from the explicitly stored memory. However, this approach suffers from scalability issues as the number of task increases. The second class of approaches, i.e., regularization approaches, overcome catastrophic interference by imposing constrains on the neural network parameters. [175] proposed learning without forgetting algorithm that enforced regularization via knowledge distillation. Specifically, given a set of shared parameters across all tasks, it optimizes the parameters of the new task together with the shared parameters. [140] proposed to incrementally train the autoencoder by adding neurons for training samples with high loss. Similarly, [176] proposed an algorithm that incrementally trains the network to grows in a hierarchical fashion. More recently, [116] proposed a dynamically expanding network that increases the number of parameters to incrementally train the models on new tasks. Concretely, it performs selective retraining that expands the network capacity using group sparse regularization.

While the aforementioned approaches made significant advances, they mainly focused on alleviating the issue of catastrophic forgetting by designing solutions that operate over the parameter space. Different from these, the proposed research alleviates the issue of catastrophic forgetting by designing a new regularization technique that operates over the embedding space rather than the parameter space.

7.3 Approach

Our goal is to continually integrate multiple kinds of KBs into the pretrained biomedical language models, and generate knowledge-powered representations. To achieve this, we develop a new representation learning approach that first models the hierarchical structure of KBs (Section 7.3.1), and then proposes a continual learning scheme to integrate multiple KBs in a perpetual manner (Section 7.3.2). Formally, let us consider an overall vocabulary $\mathcal{V} = \{w_1, \ldots, w_V\}$ of size V. $\mathbf{w}_i \in \mathbb{R}^d$ denotes a $d-$dimensional word embeddings that can be derived from any pretrained language representation models, e.g., BioBERT [2] or PubMedBERT [138]. Given a pretrained set of word representations $(\mathbf{w}_1, ..., \mathbf{w}_V)$, the objective is to update the representations for each word $w_i \in \mathcal{V}$ by infusing the semantic knowledge from the successive KBs, i.e., $(\mathbf{w}_1^1, ..., \mathbf{w}_V^1), (\mathbf{w}_1^2, ..., \mathbf{w}_V^2), \ldots, (\mathbf{w}_1^N, ..., \mathbf{w}_V^N)$ for unbounded N, where N refers to the n-th KB. Figure 7.2 shows an overview of the proposed approach.

7.3.1 Modeling Hierarchical Knowledge-base

A hierarchical KB represents a directed acyclic graph $\mathcal{G}$ with two intrinsic topological parameters, i.e., ancestors and siblings. Ancestors refer to the direct and indirect hypernyms of a concept, whereas siblings refer to the set of concepts located at the same level. Modeling the complementary information from both the ancestors and siblings can enable us to capture the taxonomic structure of KBs in a comprehensive manner.
Modeling Ancestors: The meaning of a concept is formed by the accumulation of the features coming from a higher ancestor to another less deep. Thus, we propose to quantify the contribution of each ancestor, i.e., both direct and indirect parents of a concept. This is different from the existing approaches lauscher2019informing that only consider modeling the direct parents. Specifically, each concept w_i in $\mathcal{G}$ is assigned a basic embedding vector $\mathbf{e}_i \in \mathbb{R}^d$. Next, we formulate a concept's final representation $\mathbf{m}_i$ as a

convex combination of the embeddings of itself and its ancestors:

$$\mathbf{m}_i = \sum_{j \in \mathcal{N}(i)} \beta_{ij} \mathbf{e}_j, \sum_{j \in \mathcal{N}(i)} \beta_{ij} = 1, \beta_{ij} \geq 0 \text{ for } j \in \mathcal{N}(i) \tag{7.1}$$

where $\mathbf{m}_i \in \mathbb{R}^d$ denotes the final representation of the concept w_i, $\mathcal{N}(i)$ denotes the indices of the concept w_i and its ancestors, $\mathbf{e}_j$ the embeddings of the concept w_j, and $\beta_{ij} \in \mathbb{R}^+$ the attention weight on the embedding $\mathbf{e}_j$ when computing $\mathbf{m}_i$. The attention weight β_{ij} in Equation 7.1 is calculated by the following softmax function:

$$\beta_{ij} = \frac{exp(f(\mathbf{e}_i, \mathbf{e}_j))}{\sum_{k \in \mathcal{N}(i)} exp(f(\mathbf{e}_i, \mathbf{e}_k))} \tag{7.2}$$

where $f(\mathbf{e}_i, \mathbf{e}_j)$ calculates the compatibility between the basic embeddings of $\mathbf{e}_i$ and $\mathbf{e}_j$ via a scoring function. Specifically, the scoring function is approximated by a single layer perceptron:

$$f(\mathbf{e}_i, \mathbf{e}_j) = \mathbf{S}^T \tanh(\mathbf{Q} \begin{bmatrix} \mathbf{e}_i \\ \mathbf{e}_j \end{bmatrix} + \mathbf{b}_1) \tag{7.3}$$

where **S, Q** and $\mathbf{b}_1$ are the parameters to be learned. *tanh* is the activation function of the hidden layer. In order to learn these attention weights, we propose to train the model on a multi-label classification task where the objective is to predict the labels for the biomedical articles. Existing studies [177] have found this to be an effective strategy for learning high-quality predictive embeddings. To train the model on the multi-label classification task, we use another single layer perceptron:

$$\hat{y}_j = \text{softmax}(\tanh(\mathbf{P}\mathbf{m}_i + \mathbf{b}_2)) \tag{7.4}$$

where $\mathbf{P}$ and $\mathbf{b}_2$ are the learnable parameters. Finally, we use the cross-entropy loss as the objective function for the predictive task as follows:

$$\mathcal{L}_c = -\frac{1}{NK} \sum_{i=1}^{N} \sum_{j=1}^{K} y_{ij} log(\hat{y}_{ij}) + (1 - y_{ij}) log(1 - \hat{y}_{ij}) \tag{7.5}$$

where N refers to the number of articles, and K refers to the number of labels. $\hat{y}_{ij}$ and y_{ij} refers to the predicted probability and true value, respectively, for the i-th article and the j-th label. Using the above mechanism, we model the semantic contribution of each ancestor on the concept's final representation in a comprehensive manner.

Modeling Siblings: Concepts that are descendants of a common parent are referred to as siblings. Siblings possess both the common and distinctive semantics with respect to one another. While modeling the ancestors accounts for the semantic commonality between siblings, their specific semantic differences remain ignored. For example, in Figure 7.1, the concepts 'Heart Failure, Diastolic' and 'Heart Failure, Systolic' are close to 'Heart Failure' as they inherit the same attribute. However, they should be differentiated from each other as they also hold significantly different attributes. Unlike the

existing approaches [165, 167] that mainly exploit the hyponyms-hypernyms relationships, we argue that modeling the discriminative semantics between siblings can assist in capturing the semantic richness of taxonomies in a coarse-grained manner. Technically, this is achieved by widening the semantic distance between the embeddings of concepts at the same level. The training objective to minimize is the following:

$$\mathcal{L}_s = \sum_{j \in Siblings(w_i)} cos(\mathbf{e}_i, \mathbf{e}_j); i \neq j \tag{7.6}$$

where $Siblings(w_i)$ refers to the set of siblings for concept w_i. $cos(,)$ denotes the similarity measure function computed by:

$$cos(\mathbf{e}_i, \mathbf{e}_j) = \frac{\mathbf{e}_i^T.\mathbf{e}_j}{|\mathbf{e}_i| \cdot |\mathbf{e}_j|} \tag{7.7}$$

By combining the objective functions of both the ancestors and siblings, we can derive the overall objective function as $\mathcal{L} = \mathcal{L}_c + \lambda_1 \mathcal{L}_s$. λ_1 is the balancing hyperparameter. The pseudocode of the proposed knowledge modeling technique is shown in Algorithm 1.

7.3.2 Continual Knowledge Infusion

In this section, we describe our efforts to continually integrate the KBs into the pre-trained language models. Let $\{KB_1, \ldots, KB_{n-1}, KB_n\}$ and $\{\theta_1, \ldots, \theta_{n-1}, \theta_n\}$ denote a set of existing KBs and their embedding representations respectively. Given the embedding representation θ_1 generated from KB_1 (via modeling the KBs ancestors and siblings), we propose to incrementally fuse the successive KBs by initializing the embeddings $\boldsymbol{\theta_n}$ of KB_n with $\boldsymbol{\theta_{n-1}}$ of KB_{n-1}. This initialization scheme aligns the learned embeddings in the unified coordinate space [178] and enables continual knowledge infusion by performing direct knowledge transfer. While this straightforward mechanism works well in practice, it might lead to catastrophic forgetting [129] when a large number of KBs need to be integrated. Concretely, as we train the model on new KBs, the embedding space might become distorted and thus forget the previously acquired knowledge. Some existing continual learning approaches [174, 171, 134] have attempted to alleviate this issue (i.e, catastrophic forgetting). However, they mainly operate over the models parameter space which is different than the current setting of embedding space. Additionally, the existing approaches are mainly proposed for the multi-task settings whereas the current setting is single-task incremental (i.e., progressively integrating diverse KBs). To address these challenges, we propose a new regularization mechanism that mitigates the issue of catastrophic forgetting in the embedding space

Algorithm 2 Learning knowledge-powered representations

Input: Hierarchical knowledge-base G, dimensionality d of the word embeddings, word vocabulary $\mathcal{V}$
Output: Knowledge powered embeddings $\mathbf{M}$
repeat
  for concept c_i **in** $\mathcal{V}$ **do**
    if c_i **in** G **then**
      Refer G to find c_i's ancestors C'
      for concept c_j **in** C' **do**
        Calculate attention weight β_{ij} using Equation 2
      end for
    end if
  end for
  Calculate prediction loss $\mathcal{L}_c$ for ancestors using Equation 5
  Calculate prediction loss $\mathcal{L}_s$ for siblings using Equation 6
  Obtain the final representations $\mathbf{M}$ using $\mathcal{L} = \mathcal{L}_c + \lambda_1 \mathcal{L}_s$
  Update the parameters according to the gradient of $\mathcal{L}$
until converge

for the single-task incremental settings. Simply put, an effective regularization mechanism should allow the embedding updates to accommodate the new knowledge and at the same time ensure the changes do not loose the previously acquired knowledge. To achieve this, we propose to identify the concepts whose embeddings changed sporadically (i.e., unstable) during the successive KB integration. To measure the embedding stability, we propose to quantify the portion of overlapping words between the concept's k-nearest neighbors from their successive embedding spaces. Concretely, given a concept w_i, let c_1 and c_2 be the k-nearest neighbors of w_i in the consecutive embedding spaces KB_{n-1} and KB_n. The stability value for concept w_i is the ratio of overlapping words in c_1 and c_2. All concepts with the stability value below a pre-defined threshold are deemed as unstable.

Having identified the unstable concepts, we propose to minimize the variance of representations in the embedding space. Since the embeddings are in the unified coordinate space, we argue that minimizing the variance of concepts representations learned over successive KBs can mitigate the issue of catastrophic forgetting. Specifically, given two

neighbourhood sets c_1 and c_2 for a concept w_i, the difference of representations in the successive embedding space can be defined by the L_2 norm,

$$d_{c_1,c_2}(w_i) = ||(\mathbf{m}_i)_{c1} - (\mathbf{m}_i)_{c2}||_2 \tag{7.8}$$

Let B ($B \subseteq \mathcal{V}$) denote the set of concepts that have significantly distorted representations in the successive embedding spaces. We propose to minimize the variance in their representations using the hinge loss:

$$L_H = \sum_{w_i \in B \text{ and } w_i \in c_1 \cap c_2} [d_{c_1,c_2}(\mathbf{X}\mathbf{m}_i) + \gamma - d_{c_1,\hat{c}_2}(\mathbf{X}\mathbf{m}_i)]_+ \tag{7.9}$$

where $\mathbf{X} \in \mathbb{R}^{d \times d}$ is a transformation matrix and $\gamma > 0$ is a hyper-parameter that denotes the margin. To enlarge the semantic distance with unrelated neighbours, we generate the negative samples $(c_1, \hat{c}_2)$ by substituting c_2 with a random neighbour set $\hat{c}_2$. The operator $[x]_+$ denotes $max(x, 0)$. We impose an orthogonal regularization on $\mathbf{X}$ to prevent the information loss.

$$L_0 = ||\mathbf{I} - \mathbf{X}^T\mathbf{X}||_F \tag{7.10}$$

where $||.||_F$ denotes the Frobenius norm, and $\mathbf{I}$ is an identity matrix. The overall learning objective of the proposed regularization scheme is then denoted as $L = L_H + \lambda_2 L_O$ with a positive hyperparameter λ_2. In this way, the proposed approach prevents the issue of catastrophic forgetting by minimizing the variance of concept representations learned over successive KBs. It is worth nothing that the above formulation (Equation 7.8, 7.9) apart from mitigating catastrophic forgetting also models the interplay between multiple KBs that results in knowledge enriched concept representations. Finally, we concatenate the produced knowledge representations with the pre-trained language representations to generate knowledge-powered representations that can be utilized as the input feature for the task-specific layer of the downstream tasks (refer Figure 7.2).

7.4 Experiments

7.4.1 Datasets

- Named entity recognition (NER): For the biomedical NER, we choose BioCreative II Gene Mention (BC2GM) [151] and Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA) [179] as our datasets. The BC2GM dataset contains sentences from PubMed that are annotated with gene entities. We use the pre-processed set of train (15197), development (3061),

and test (6325) splits released by [138] for our experiments. The JNLPBA dataset is another NER corpus that is annotated with entities such as protein, DNA, RNA, cell line, and cell type. Similar to the BC2GM dataset, we use the pre-processed set of train (46750), development (4551), and test (8662) splits released by [138].

- Relationship Extraction (RE): For the biomedical RE, we utilize the widely used datasets such as CHEMPROT [154] and GAD [180] for our experiments. These datasets contain protein-chemical and gene-disease relations respectively. The total number of training/dev/test splits for CHEMPROT and GAD are 18035/11268/15745 and 4261/535/534 respectively. The preprocessed datasets are available at [2].

- Question Answering (QA): For the biomedical QA task, we use BioASQ 7b-factoid [181] and BioASQ 6b-factoid [182] as our datasets. These datasets contain factoid question answers that are annotated by the biomedical experts. Since the baseline algorithms choose the factoid part of BioASQ datasets for their experiments, we followed the same practice. The total number of train/test for BioASQ 7b-factoid and BioASQ 6b-factoid are 670/140 and 618/161 respectively. We use the preprocessed datasets released by [2] for our experiments.

Hierarchical Knowledge-Bases

We choose to integrate three well-known hierarchical knowledge-bases, i.e., Medical Subject Headings (MeSH) [145], Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) [170], and International Classification of Diseases (ICD-10) into the pretrained contextualized language models. MeSH is a taxonomic resource wherein the concepts are arranged in a hypernym-hyponym relationships. The concepts are organized into 16 sub-trees such as Anatomy, Organism, Diseases and so on. SNOMED-CT is another parent-child hierarchy that contains more than 300,000 concepts. These concepts are organized into 18 sub-trees such as clinical findings, procedures, organisms, and so on. ICD-10 is an ontological resource that contains over 70,000 concepts. Similar to the MeSH and SNOMED-CT, the concepts in ICD-10 are also organized into a hierarchical structure.

Table 7.1: Comparison of prediction performance and training efficiency in the bioNLP datasets. The evaluation metric for BC2GM, JNLPBA, CHEMPROT, and GAD is micro-F1. Accuracy is reported for BioASQ 7b-Factoid and BioASQ 6b-Factoid. To measure training efficiency, we report FLOPS.

Model	BC2GM	JNLPBA	CHEMPROT	GAD	BioASQ 7b-Factoid	BioASQ 6b-Factoid	#FLO
BioBERT	80.113	76.223	70.023	75.228	80.872	71.782	N/A
BioBERT-KB (**Proposed**)	84.189	79.554	74.229	79.389	82.982	74.897	N/A
SciBERT	79.721	75.112	70.113	74.112	76.762	71.998	N/A
SciBERT-KB (**Proposed**)	84.221	79.943	74.922	78.343	78.223	75.218	N/A
PubMedBERT	80.982	75.112	71.121	76.298	82.872	70.221	N/A
PubMedBERT-KB (**Proposed**)	84.287	79.742	74.912	80.421	84.872	74.264	N/A
BioELMo	81.198	75.111	69.123	75.998	73.982	72.221	N/A
BioELMo-KB (**Proposed**)	85.932	79.432	74.299	79.732	75.845	76.223	N/A
LIBERT	83.223	79.845	73.521	78.193	87.223	73.955	N/A
LIBERT-KB (**Proposed**)	86.873	83.392	76.117	81.231	88.929	76.892	N/A
KnowBERT	82.114	78.984	72.112	78.421	87.111	72.198	N/A
KnowBERT-KB (**Proposed**)	85.821	81.932	75.367	81.754	88.228	76.823	N/A
SenseBERT	81.432	77.231	71.122	77.289	86.121	72.143	N/A
SenseBERT-KB (**Proposed**)	84.733	80.744	74.763	79.833	87.276	75.734	N/A
BERT-MK	81.832	77.872	72.843	77.341	86.222	72.397	N/A
BERT-MK-KB (**Proposed**)	84.763	80.733	75.234	80.721	87.989	76.633	N/A
K-BERT	80.231	75.245	70.123	74.908	85.227	71.983	N/A
K-BERT-KB (**Proposed**)	83.145	78.932	73.982	77.871	86.278	73.172	N/A
BIOBERT-LWF	81.632	77.672	71.812	76.891	81.672	72.984	19.2E
BIOBERT-MAS	82.023	78.945	73.619	77.904	83.872	73.764	17.2E
BIOBERT-EWC	83.192	79.893	73.892	78.009	83.989	74.167	15.3E
BIOBERT-Continual (**Proposed**)	86.732	83.873	76.981	82.982	85.321	77.983	9.4B

Pre-training and Task-specific Settings

Pre-training: We train the state-of-the-art (SOTA) contextualized language models such as BioBERT [2], SciBERT [125], PubMedBERT [138], and BioELMo [1] on the same training corpus. We choose the latest collection of PubMed[1] abstracts as our pre-training corpus. The corpus contains 14 million abstracts that are predominantly from the areas of life sciences and biomedicine. Following suggestions from the previous studies [2, 173], we utilize the Natural Language Toolkit (NLTK) [183] to split sentences. The sentences that are less than 5 words are removed. Altogether, a large corpus containing 3.2 billion words is achieved for the language model pre-training. We use the publicly released source codes of these language models to conduct the pre-training procedure. Since our goal is to perform effective knowledge infusion, the concepts appearing in the corpus needs to be aligned with the concepts present in the hierarchical KBs. To achieve this, we utilize the concept normalization algorithm proposed in [184]. The algorithm[2] maps the concept mentions in the natural language text to their corresponding concept entries in the standardized biomedical thesaurus, i.e., Unified Medical Language Systems (UMLS) [185]. Generally speaking, the approach first applies a candidate generator to generate a list of candidate concepts, and then use a BERT [142]-based list-wise classifier to rank the candidate concepts. The main advantage of this approach is that it considers both the morphological and semantic information to perform accurate concept normalization. Experiments conducted on three social media datasets, TwADR-L [186], SMM4H-17 [187], AskAPatient [186], and one clinical notes dataset, MCN [188] validate the efficacy of the proposed approach. TwADR-L and SMM4H-17 contains 5,074 and 9,149 adverse drug reaction (ADR) annotations that are mapped to 2,220 and 513 concepts from the Medical Dictionary Regulatory Activities respectively. Similarly, AskAPatient and MCN contains 17,324 and 13,609 concept mentions that are mapped to 1,036 and 3,792 concepts from the SNOMED-CT respectively. For all these four datasets, the performance is measured using accuracy metrics averaged over the 10-fold cross validation. The results are compared with the existing concept normalization algorithms [189, 190]. Overall, the chosen concept normalization algorithm [184] obtains SOTA performance in two social media datasets and one the clinical dataset. Specifically, the impact of performance (in terms of accuracy gains) is highest in the clinical domain, i.e., MCN dataset (83.56%), whereas the performance is lowest in one

[1] ftp://ftp.ncbi.nlm.nih.gov/pubmed/baseline
[2] https://github.com/ dongfang91/Generate-and-Rank-ConNorm.

of the social media dataset SMM4H-17 (Achieved: 88.24%, SOTA: 89.64%).

Task-specific Settings: All of the chosen downstream bioNLP tasks, i.e, named entity recognition, relationship extraction and question answering can be formulated as a classification problem. Consequently, we fine-tune the models by adding a single linear layer on top of each contextualized embedding models and then train them using the task-specific training data. Following the practice in the existing literature [2, 125], the range of hyperparameters such as learning rate, batch size, and epoch number is chosen within the range of [1e-5, 3e-5, 5e-5], [10, 16, 32, 64], and [2–60] respectively. After analyzing the average prediction performance in the validation sets, the learning rate, batch size, and epochs are set to 3e-5, 16 and 6 respectively.

Baseline Models

(1) The first class of baselines are the state-of-the-art (SOTA) contextualized language models [2, 125, 138] that are trained solely on the biomedical corpus. Our main objective in this experiment is to measure the boost in performance achieved by these models when integrated with multiple kinds of KBs. (2) The second class of baselines are the KB-augmented methods [165, 167] that learn concept representations by injecting the external semantic knowledge into the pretrained contextualized language models. Our objective is to compare the performance with the existing knowledge-powered approaches. Note that we do not compare our results with the KB-augmented approaches [191, 173] that are not designed for integrating the hierarchical KBs. (3) The third class of baselines are the continual learning approaches [129, 175] that can be adapted to the current NLP setting. Our objective is to compare the performance of the proposed continual learning strategy with the existing continual learning approaches. We choose these approaches [129, 175] as our baselines because they can be adapted to the continual learning settings for NLP tasks.

7.4.2 Results and Discussion

Table 7.1 reports the results on the tasks of named entity recognition, relationship extraction, and question answering. Following the evaluation criteria used in the previous studies [2, 125, 138], we report the micro F_1 score for named-entity-recognition, relationship extraction and accuracy for question-answering. We utilize the same evaluation metrics as the baseline algorithms to facilitate a standardized comparison of

results. From the results, we make the following observations:
(1) All of the SOTA contextualized language models (without knowledge infusion) achieve competitive performance on the bioNLP tasks. No specific model significantly outperforms the others. This is reasonable because all of the aforementioned models adopt the same transformer [142] based architecture at its core. Some existing studies [138] have reported higher performance gains for the PubMedBERT model. However, this result has been attributed to the choice of the training corpus. As our objective in this research is to quantify the gains achieved from integrating KBs, the training corpus for all the SOTA models are fixed (refer Section 7.4.1). We observe that the proposed knowledge-powered versions of the SOTA models obtains significant improvement in performance. This result demonstrates that the semantic knowledge from KBs play a positive role in improving the feature representations of concepts. Analyzing the results, we observe that the incorporation of prior knowledge is especially useful for concepts with paucity of co-occurrence information in the training corpus, i.e., rare or domain-specific concepts. For instance, the concept *Myocarditis* is rarely observed in the training corpus, and thus its semantic proximity with related concepts such as *Heart* is relatively low, i.e, cosine distance = 0.39. However, their semantic relatedness is well captured via the MeSH hierarchy. As such, after integrating the MeSH hierarchy into the language models, the semantic proximity between their feature representations is significantly boosted, i.e, cosine distance = 0.67. This result indicates that the proposed approach can learn robust representations for the rare (or domain specific) concepts. To further evaluate our approach, we perform case-studies on two rare biomedical concepts (i.e., Peritonitis and Atherosclerosis) [124, 115]. For both of these concepts, we analyze their top 5 nearest neighbour returned by the baseline algorithm (BioBERT) and the proposed approach (BioBERT-KB). The top 5 neighbours for 'Peritonitis' returned by BioBERT and BioBERT-KB are ['Gastroenteritis','Empyema','Enterocolitis','Pyomyositis','Esophageal
Diseases'] and ['Peritoneal Fibrosis','Peritoneal Neoplasms','Peritonitis, Tuberculous','Panniculitis, Peritoneal','Pneumoperitoneum'] respectively. Similarly, the top 5 neighbours for 'Atherosclerosis' returned by BIOBERT and BioBERT-KB are ['Fibromuscular Dysplasia','Angioedema','Angio matosis','Angiodysplasia',' Moyamoya Disease '] and ['Arteriosclerosis Obliterans','Peripheral Arterial Disease','Coronary Artery Disease','Intra
cranial Arteriosclerosis','Arterial Occlusive Diseases'] respectively. As it can be observed, the neighbours returned by the proposed approach, i.e., BioBERT-KB, forms

more meaningful semantic coherence than those returned by the baseline algorithm, i.e., BioBERT. This result indicates that the injection of prior knowledge (proposed approach) helps to improve the representations of rare or domain-specific words.

(2) Analyzing the results with the second-class of baselines, i.e., knowledge-augmented algorithms such as [165, 167, 166, 173], we observe that most of the knowledge-powered models perform better than the purely data-driven models. This result reinforces the usefulness of exploiting the semantic knowledge. For a head-to-head comparison, we compare the proposed approach with the existing approaches [173, 192] that also focus on integrating the medical knowledge graph (KG). BERT-MK [173] models the subgraphs in the medical KG and injects the graph contextualized knowledge into the pretrained language model, whereas K-BERT [192] injects the domain knowledge in the form of semantic triples (e.g., 'Diabetes Mellitus, Type 1', *Child-of*, 'Diabetes Mellitus'). As it can be observed from the Table 7.1, the proposed approach outperforms BERT-MK and K-BERT in all the bioNLP tasks. This result indicates that whilst integrating the semantic knowledge can boost the performance, the methods to integrate the structure of knowledge has direct implications on the overall performance. Moreover, we also compare the performance of our approach with existing approaches such as LIBERT-lauscher2019informing and KnowBERT [167] that are also designed for integrating the hierarchical structure of KBs. However, they do not perform on par with the proposed approach. We speculate this is because whilst the existing knowledge-augmented algorithms incorporate the ancestral information present in hierarchical KBs, they usually ignore to model the *discriminative* semantic information present between concept's siblings. This is limiting because the "siblings" contribute valuable semantic information too (refer Table 7.2). The proposed approach models both the ancestral and sibling relationship present in hierarchical KBs, and the overall results show that this strategy is both reasonable and effective. We note that the proposed approach obtain SOTA results within the realm of pre-trained biomedical language models. Some existing approaches [193] that also explore the usage of biomedical KBs (in different problem setting) have reported approximately 5% improvement in F-score for the same relationship extraction task. This result indicates that the other (or more accurate) methods to exploit KBs may result in better performance for the same bioNLP tasks, and thus further comparison studies should be conducted to examine this issue.

(3) Comparison with the existing continual learning baselines is performed to examine whether the proposed approach can effectively mitigate the issue of catastrophic

forgetting. Following the practice in the existing literature, we report average accuracy [194] obtained after incrementally integrating all three KBs. As it can be observed from the Table 7.1, EWC appears to be the strongest baseline. The performance of MAS is marginally better than LWF. We believe this is because EWC does not overly constrain the embeddings and thus has better knowledge retention capabilities. The proposed continual learning technique performs significantly better than the existing approaches. We speculate two reasons for this: (a) Most of the existing continual learning approaches [194] are designed for computer vision tasks, and thus their direct application to the realistic NLP tasks yields unsatisfactory results [195]. This is due to the fundamental differences in the properties of imaging and textual datasets. (b) Existing continual learning approaches mainly operate over the neural networks' parameter space. However, deep neural networks are known to posses a huge number of parameters, and thus the methods operating over the parameter space can be computationally expensive [171]. As such, the ability of these approaches to overcome the issue of catastrophic forgetting is limited. To address this, we formulate the continual learning problem in the embedding space (as opposed to the parameter space). This strategy is both memory efficient and at the same time has better knowledge retention capabilities. Consequently, the issue of catastrophic forgetting is effectively mitigated (refer Table 7.1 - third block). Moreover, we also measure the computational efficiency of the proposed approach with the existing continual learning approaches. Note that for a head-to-head comparison, we report the computational efficiency with the continual learning baselines only. Specifically, for measuring the computational efficiency we report floating-point operations (FLOPs) [127]. FLOPs calculate the number of floating-point operations that the models perform for a single process. From the results in Table 7.1, we can observe that the proposed approach effectively preserves the memory efficiency of the contextualized language models. We speculate two reasons for this: (a) Different from the existing approaches that require a working memory to store the informative samples from previous tasks/snapshots, the proposed approach directly operates over the embedding space. (b) Since the proposed regularization mechanism selectively identifies the concepts that require an embedding update, it is more efficient than approaches [171] that update the embeddings in a batch-mode fashion.

Analyzing contribution of ancestors/siblings and individual KBs

To quantify the semantic contributions from both the ancestors and siblings, we develop variants of the SOTA models, e.g, $BioBERT_{Ancestors}$, $BioBERT_{Siblings}$, and generate representations. We choose one dataset from each of the three bioNLP tasks and report the results in Table 7.2. As it can be observed, modeling both the ancestors and siblings contribute uniquely towards capturing the topological properties of the hierarchical KBs. The accuracy gain margins are higher for the ancestors. We speculate this is because the cardinality set of ancestors is greater than the siblings. Nevertheless, the siblings provide distinctive semantic information too. This can be observed from the result that the contribution of "Ancestors+Siblings" obtains the best result. In summary, the results validate the importance of incorporating both the ancestors and siblings to preserve the comprehensive structure of hierarchical KBs.

In another ablation study, we study the benefits of integrating multiple hierarchical KBs. Table 7.3 reports the results. All of the ablated versions use BioELMo [1] as the backbone model. It can be observed that the KBs (i.e., SNOMED-CT and MeSH) contribute more accuracy gains to the overall performance. We believe this is due to the broader coverage of biomedical concepts in the SNOMED-CT/MeSH. The combination of multiple KBs achieves the best performance. We also conducted experiments whilst shuffling the order of KBs. However, we did not observe any noticeable change in the results. Overall, the result suggests the practical benefits of integrating multiple kinds of KBs into the biomedical language models. More importantly, it demonstrates the necessity of designing new continual learning approaches that can integrate diverse KBs in a progressive manner.

7.4.3 Hyper-Parameter Settings

The experiments are conducted on NVIDIA TITAN Xp GPU. For a head-to-head comparison, we set the size of embeddings generated by both the proposed approach and the baseline algorithms to 200. The hyper-parameters for the loss functions are tuned using grid search on the validation set. Following the convention in existing studies [2, 138], we set the range to [0.01, 0.05, 0.1, 1, 10]. The best performance for λ_1 and λ_2 is obtained at 0.05 and 0.1 respectively.

Table 7.2: Analyzing the semantic contribution of ancestors and siblings using SOTA biomedical language models.

Models	BC2GM	CHEMPROT	BioASQ
BioBERT$_{Ancestors}$	82.783	72.123	78.198
BioBERT$_{Siblings}$	80.432	70.751	76.763
SciBERT$_{Ancestors}$	82.456	71.219	75.638
SciBERT$_{Siblings}$	80.912	70.094	72.281
BioELMo$_{Ancestors}$	82.562	72.945	70.229
BioELMo$_{Siblings}$	80.903	70.091	68.398
BioBERT$_{Ancestors+Siblings}$	84.189	74.229	80.872
SciBERT$_{Ancestors+Siblings}$	84.221	74.992	77.762
BioELMo$_{Ancestors+Siblings}$	85.932	74.299	73.982

Table 7.3: Analyzing the semantic contribution of individual KBs using BioELMo [1]

KBs	BC2GM	CHEMPROT	BioASQ
MeSH	81.093	70.566	70.227
SNOMED-CT	81.234	70.984	70.981
ICD-10	81.013	70.094	68.241
MeSH + ICD-10	83.093	72.187	71.102
SNOMED-CT + ICD-10	83.112	72.094	71.011
MeSH + SNOMED-CT	83.903	72.987	71.912
MeSH + SNOMED-CT + ICD-10	85.932	74.299	73.982

7.5 Conclusions

In this research, we proposed a new representation learning approach that continually infuses the semantic knowledge from the hierarchical KBs into the pretrained biomedical language models. Specifically, the approach models the unique topological properties, i.e., ancestors/siblings of the hierarchical KBs and efficiently updates the concept representations whilst integrating diverse hierarchical KBs. Overall, the proposed approach generates high-quality knowledge-powered representations and at the same time preserves/improves the learning efficiency of SOTA contextualized language models. Comprehensive experiments conducted on the bioNLP tasks validate the efficacy of the proposed approach, and suggests that the proposed framework is capable of generating meaningful representations that are useful for a variety of biomedical applications.

Part III

Hypothesis Generation For Accelerating Scientific Discovery

Chapter 8

Uncovering Conceptual Bridges Based on Concept Evolution

8.1 Introduction

Scientific knowledge is growing at an unprecedented rate as evident from the growing body of research publications, grants, clinical trials and other scientific endeavors. A large body of this knowledge available in the free-form text has provided practitioners access to a staggering amount of information; however, at the same time, it has also made it increasingly difficult for them to keep up with the latest information, trends and findings in their field of interest in a reasonable amount of time. Imagine a researcher attempting to formulate a new hypothesis in the research area of *autism* (a serious developmental disorder). To do so, first, one has to thoroughly study and understand the existing body of literature already available. At present, a simple search in MEDLINE (a popular bibliographical database) for autism yields more than 50,000 results. While technologies based on text summarization would help users get a high level idea of the papers, it fails to stitch together disparate and seemingly uncorrelated facts together to present novel and "actionable" insights that can drive new research frontiers. Motivated by this, hypotheses generation, a sub-branch of biomedical text mining, aims at identifying non-trivial implicit assertions within a large body of documents. Simply put, the task of hypotheses generation is to answer questions like: Is there an implicit linkage between two seemingly related but explicitly disjoint topics of interest (A and

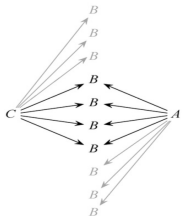


Figure 8.1: An Overview Schematic of Hypotheses generation

C)? Consider the example shown in Figure 8.1. It can be observed that a direct relationship between two topics A and C might not be known/studied but there might exist an implicit linkage between them via bridging terms (B). Finding these *conceptual bridges* might reveal hitherto unknown but potentially interesting relationships. This is the crux of the problem that this paper attempts to address.

Prior studies tackle this problem through a range of solutions based on approaches such as distributional statistics [9, 196], graph theoretic measures [10, 11] and supervised machine learning techniques [12, 197]. However, in a broad sense, they are afflicted with three major drawbacks:

1. **Rigid schema**: Almost all of the previous approaches rely on a "hard-wired" schema (e.g. graph) that results in finding only those linkages that are en route. Consequently, it risks missing the connections that are surprising or radical. More often, these radical linkages have the potential of shedding novel insights into pathways that would remain otherwise hidden.
2. **Strict query reliance**: Existing approaches find implicit connections by strictly relying on the given input pairs; thereby ignoring the subtle cues from concepts present in their local neighbourhood.
3. **Static domain**: The prior studies mainly assume the prevailing domain to be static; nevertheless, it is known that the domains in general (and in particular bio-medicine) are usually dynamic with new facts being added every single day [13].

To tackle the problem of **rigid schema**, we model the problem of finding key conceptual bridges in the latent continuous space which allows us to include even those terms in our search-space that have not yet been rigorously investigated; thereby nudging the system to perform novel and radical discovery. We use the concept of word embedding techniques [18, 66] in conjunction with temporal information to identify bridge terms that have the highest likelihood of creating a meaningful connection.

The use of word-embeddings also allows us to circumvent the second issue of **strict query reliance**. Because word-embeddings project semantically similar terms closer in the vector space [18], we can leverage the terms that are deemed 'close enough' to the query to augment our search-space. This differs markedly from the classical approaches [10, 198] that find conceptual bridges by relying solely on the user provided input query terms. Their idea being that those concepts that have high semantic relatedness to both the start and end concepts (A and C) are promising candidates for bridge concepts. In this study, we extend this intuition and argue that good bridge concepts are those that, apart from being connected to A and C, are also connected to their semantically similar neighbours.

The infusion of temporal information into our word-embedding generation process enables the proposed model to be sensitive to the dynamic nature of the domain; thus alleviating the limitations of modeling it under **static domain**. While some prior studies [67, 66] have attempted to generate temporally sensitive word-embeddings, they cannot handle the current problem setting, wherein it is important for the temporal embeddings to factor in the fundamental relationship between input query and its informative local context in order to find promising conceptual bridges. To this end, we propose a new approach that allows us to first train the distributed representation of words in temporally distant time scopes and then learn a mapping function/transformation matrix being sensitive to both the global and query-specific semantics; thereby enabling the system to learn precise transformation.

Thus, our contributions can be summarized as:

1. We propose a novel model for hypotheses generation, namely *Concepts-Bridges*, that infers implicit relations by capturing the latent evidence manifested in the temporal drift.
2. The proposed technique for capturing temporal dynamics is sensitive to both local and global correspondence of input query, thereby capturing the semantics at a granular level.
3. The experimental results corroborate the efficacy of the proposed model - we obtain a 20% improvement over baselines in terms of Mean Average Precision @ top-K. Qualitative evaluation of the bridge terms also validate that the hypotheses generated are plausible and worthy of further investigations.

8.2 Related Work

Hypothesis generation from unstructured text has long been an important problem of text mining [198, 199]. This area of study in particular started gaining attention after the seminal work of Don R. Swanson in 1986 [5]. In this study, the researchers demonstrated the potential of combining facts from multiple documents to discover new knowledge. However, their approach required significant manual labor. To overcome this issue, the subsequent studies focused on automating it.

Distributional approaches: Some of the previous studies in this area of research relied on statistical analysis of concept co-occurrence (term frequency, inverse document frequency, record frequency and so on) [9, 198, 196]. Their notion being, new associations are likely to be found if the conceptual bridges are highly or rarely connected to the disparate topics of interest. However, a drawback of these approaches lie in the fact that term frequencies indicate strong but not necessarily semantically meaningful associations. Another disadvantage is their neglect of temporal dimension. This is troublesome because it is known that the semantic meaning of a concept evolves over time. Furthermore, it promptly affects domain such as bio-medicine where some new facts emerge and some are rendered obsolete every now and then.

Graph theoretic approaches: Another line of research tends to model the problem of hypotheses generation using graph based approaches [11, 10, 22]. In [11], the authors proposed a graph-based approach utilizing semantic predicates present in the form of subject-verb-object. However, their performance was tied to the accuracy and coverage of such predicate extracting tools. More recently, [10] proposed a context-driven approach wherein the sub-graphs are automatically generated for the user provided input. The essence of this study was to utilize the idea of shared context to find relevant bridge concepts. While these graph based approaches have been shown more successful than distributional approaches, they still suffer from scalability issues. Moreover, as these models rely on a rigid schema, they risk missing surprising association that are not in their route. This may be limiting because one of the main objectives of hypotheses generation is to provide users with radical (but meaningful) associations.

Machine learning based techniques: Recently, several studies [12, 197] proposed supervised machine learning based approaches to generate novel hypotheses. In [12], the authors proposed a logistic regression based model to learn the characteristic path patterns of biomedical relations to infer new linkages. The machine learning

based techniques have shown the promise to find novel associations; however, a potential drawback lies in the monetary cost associated with the process of gathering training data.

Some of the motivation for this study stems from the research area of automatic language translation and temporal information retrieval [66, 23, 200]. While close in spirit, we differ from these studies in two aspects. Firstly, the goals are different. Our study focuses to capture temporal dynamics of concepts to find conceptual bridges. Secondly, our problem is more difficult in a sense that the given input is a pair of terms (instead of a single concept), and to learn accurate temporal change one has to factor in the nature of relationships between the given input pair too.

8.3 Overview of Proposed Model

In this section, we outline our proposed methodology at a high level by providing the necessary intuition behind various components in our proposed model.

Recall that the input to our system is a pair of topics of interest, which we interchangeably call as query terms. Our goal is to find temporally charged top-k bridge concepts that are most likely to connect them in future. To find these concepts in a large-scale setting, we first need a text corpus collected across time. This corpus is then split into distant time scopes to obtain the collection of articles occurring within overlapping time windows. Based on this time-specific set of articles, we extract relevant entities, represent them into the latent embedding space and then reason upon it to find novel conceptual bridges. Since the focus of this study is to capture the temporal dynamics, it is important to track the semantic evolution of concepts over time. However, due to the prevailing stochastic nature of initialization for word embedding models, a direct comparison of vector spaces to quantify bio-medical concept evolution cannot be performed. To tackle this problem, we propose to learn a transformation matrix that aligns vectors spaces across time slices and thus correspondingly encapsulates the dynamics of medical concepts. Once this alignment is performed, we can capture and rank the bridge terms by their evolving proximity to the query terms in the latent embedding space.

To learn the aforementioned transformation, we propose two ways: a) Global and b) Query-biased. While the global transformation captures the more "general" information from the corpus, the query-specific transformation captures the information particular

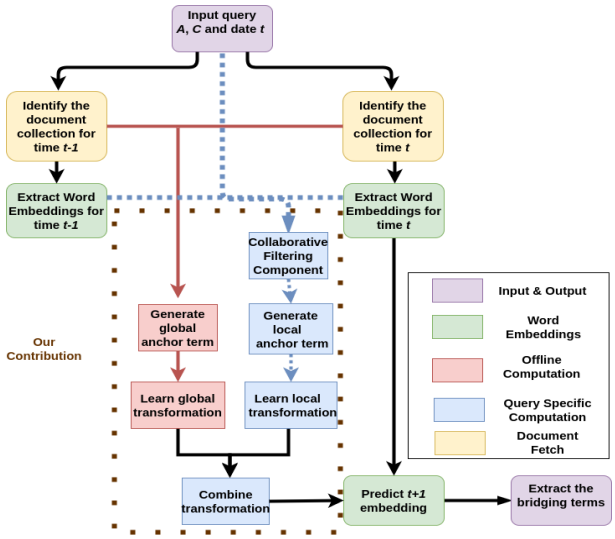


Figure 8.2: An Overview of the Proposed Framework

to the semantics of input query. To achieve the latter, we need a way to identify concepts similar to the input query so as to learn the transformation matrix utilizing them. This is where we leverage the principles of collaborative filtering. The importance of combining information from these two sources and the speculation that they complement each other is experimentally validated. Having learned the transformation matrices, we use them to calculate the likelihood of a concept to be potential conceptual bridge between the input query terms. Figure 8.2 provides a high level intuition of the proposed framework.

8.4 Methodology

This section describes our methodology in detail. It is primarily divided into three sections. Section 8.4.2 provides details on how the transformation matrix at a global level is learned from word-embeddings corresponding to the individual time-slices. Section 8.4.3 extends this idea to include the information from the local context of individual query and describes the technique to find transformation matrix in a query sensitive fashion. Having calculated both the transformation matrices, Section 8.4.4 calculates the ranked list of bridge terms.

8.4.1 Preliminaries

In this sub-section, we introduce some definitions and background information on word-embeddings.

Definition 1. Those concepts that do not change their semantic meaning over time are referred to as semantically stable concepts. An example of semantically stable concepts is "Animals". The meaning of concept "Animals" in (1850) is equivalent to its meaning in (2018).

Definition 2. Those concepts that change their meaning over time are referred to as semantically unstable concepts. An example of semantically unstable concept is "Cell". The meaning of concept cell during 1850's used to be associated with "cave", "dungeon" and "prision", however, at present (2018) it is associated with "cytoplasm", "tumor" and "epithelial cells".

Word Embeddings: To learn the distributed representation of concepts in each snapshot, we utilize a popular word embedding model, namely Continuous Bag-of-Words Model (CBOW) [18]. Given a target word w_v and its u neighboring words, the model aims at maximizing the log-likelihood of each word given its context. The objective function is shown below:

$$J = \frac{1}{V}\sum_{v=1}^{V} \log p(w_v | w_{v-u}^{v+u}) x \tag{8.1}$$

where V refers to the overall size of Vocabulary. The probability $p(w_v|w_{v+u}^{v-u})$ is calculated as:

$$\frac{exp(e_{w_v}^{\prime\top} \cdot \sum_{-u \leq j \leq u, j \neq 0} e_{w_{v+j}})}{\sum_w exp(e_w^{\prime\top} \cdot \sum_{-u \leq j \leq u, j \neq 0} e_{w_{v+j}})}$$

where e_w and e'_w denote the input and output embeddings respectively. In this study, to generate word embeddings for different time slots, we first collect all the concepts occurring in the corpus and prepare an overall vocabulary. Based on this vocabulary, we train CBOW model for each consecutive time unit. The time unit is aggregated to the granularity of ten years (e.g., 1981-1990, 1982-1991 and so on) to handle the data sparsity issue. In this setting, every concept present in the vocabulary (from the beginning of time unit) has a certain position in the vector space. Then, for each consecutive time unit, we iterate over epochs and train the word vectors until convergence. As suggested by the previous studies [18, 200], the number of embedding dimension is set to 300.

8.4.2 Global transformation

The focus of this section is to discuss the methodology of learning a global transformation matrix. This matrix is expected to capture the global temporal dynamics of concepts present in the corpus. Another objective is to align the two different vector spaces.

Having learned the distributed representation of words on distinct snapshots through Equation 8.1, the next step is to learn the transformation matrix that captures temporal change and also aligns them. In this direction, the main idea in learning a global mapping is to utilize the semantically stable terms across time as anchors to bridge the two distinct vector spaces. Once the mapping is found using anchors, other semantically unstable concepts within the two spaces can be aligned by the similarity of their positions relative to the anchor terms in their own spaces. However, this gives rise to a new challenge of selecting the candidate anchor terms. To circumvent this issue, we rely on an approximate method and choose the anchor pairs based on two criteria: a) they should have same syntactic/literal form and b) they are sufficiently frequent in both the time periods. A few examples of such terms in medical domain include "humans", "animals", "male" and so on. The rationale behind choosing frequent terms as anchors is their tendency to have high degree centrality/connectedness; this causes their position in the vector space to be semantically stable [66].

For the ease of explanation, we present the technique to learn global transformation matrix using two time stamps (t_0, t_1). Formally, given P pairs of global anchor terms (w_1^0, w_1^1),...,(w_p^0, w_p^1), where w_i^0 denotes the anchor term at time t_0 and w_i^1 denotes the anchor term at time t_1 respectively. The transformation matrix M_1 is then found by minimizing the differences between $M_1 \cdot \vec{w}_i^0$ and $\vec{w}_i^1$ (See Equation 8.2). To prevent overfitting, a regularization component is added to Equation 8.2 with γ as its corresponding weight.

$$M_1 =_{M_1} \sum_{i=1}^{P} ||M_1.\vec{w}_i^0 - \vec{w}_i^1||_2^2 + \gamma||M_1||_2^2 \tag{8.2}$$

where $\vec{w}_i^0$ and $\vec{w}_i^1$ refers to the vector position of w_i^0 and w_i^1 at t_0 and t_1 respectively. In our implementation, the top 5% frequent terms in the corpus is chosen as the size of P. Both the threshold for P and $\gamma = 0.02$ is empirically set as suggested by some of the previous studies [200].

8.4.3 Query biased transformation

The global transformation explained in Section 8.4.2 is query independent; therefore, the mappings generated are not sensitive to the specific semantics of input query. This is problematic because it generates transformation matrix that neglects the fundamental relation between query and its local context. Furthermore, this also leads to insufficient characterization of temporal dynamics that in particular affects the current problem of interest, wherein the quality of conceptual bridges is highly dependent on the given input query. To overcome this issue, we propose an approach to train the transformation matrix in a query-biased way by leveraging upon the principles of collaborative filtering. The collaborative filtering provides a systematic approach to identify terms similar to the input query terms - refer Section 8.4.3. These terms act as a "seed" to the process of generating local anchors - refer Section 8.4.3.

Generating similar concepts

Given an input concept of interest A (or C) and a date (t'), the goal is to find top-N concepts similar to the input for downstream processing. A straightforward way is to find the similar concepts by comparing the distance (in latent space) of input with each of the concepts present in the vocabulary and choosing the top-N closest neighbours. However, this becomes inefficient if the size of vocabulary scales to millions or billions. To do this in an efficient manner, we perform a soft-clustering of concepts present in the dictionary based on their word-vectors. Gaussian Mixture Model is used to perform the soft-clustering with number of clusters set to 300 as suggested in previous studies [201, 22].

Simply put, for a given input concept, we first find their respective cluster IDs and then all the concepts belonging to those clusters are added to the candidate similar set. However, this resultant set consists of concepts that are both semantically similar and semantically related to the input concept. Note that similarity calculated based on word-vectors captures both the notions of semantic similarity and relatedness [18]. This becomes problematic because in the current problem of interest we are particularly interested in finding only similar concepts. To mitigate this issue, we leverage the categorical information (known as semantic type in medical domain) of concepts. Every concept present in the vocabulary is assigned a semantic type. For example, a disease

Algorithm 3 GENERATE SIMILAR CONCEPTS

Given: Set of clusters $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$, ..., Semantic Dictionary of medical concepts *Dict*
Input: Input concept (A) and cutoff-date (t')
Output: A set (Max Heap of N) of similar concepts which are similar to a - *setSimilarConcepts*
$setSimilarConcepts \leftarrow \phi$
$\{\mathbf{C}_A\} \leftarrow clusterLookUpOn(A)$
for $x \in \mathbf{C}_{A_i}$ **do**
    $\{cand_A\} \leftarrow$ Extract all terms that have $\mathbf{C}_x$ as the cluster with highest membership probability
end for
$\{Sem_A\} \leftarrow Dict(A)$//Get the semantic type of A
for $x \in cand_A$ **do**
    $\{Sem_x\} \leftarrow Dict(x)$//Get the semantic type of x
    $CommonSem \leftarrow$Get all types of relationships existing between $\{Sem_A \times Sem_x\}$
    if $CommonSem \neq \phi$ **then**
        $setSimilarConcepts \cup \{x\}$
    end if
end for
Return *setSimilarConcepts*

such as "Migraine" is assigned to a semantic type "Disease or syndrome"[1] . We leverage this semantic information and retain only those concepts whose explicit semantic type is same as the given input. This step allows us to distill only similar concepts. Overall, this technique allows us to efficiently identify similar concepts for any given input. Algorithm 3 provides the pseudo-code for generating similar concepts.

Generating Local Anchors

Having identified a set of concepts similar to input A and C (i.e., S_a, S_c), our objective is to find set of anchor pairs, $\{¡a_1, c_1¿ \ldots ¡a_q, c_q¿\}$, such that $a_i \in S_a$ and $c_i \in S_c$. To do this, a Cartesian product between terms in S_a and S_c has to be performed. However, this

[1] The explicit semantic types of medical concepts can be obtained from Unified Medical Language System.

leads to $N \times N$ (N refers to the size of similar concept set for both A and C) comparison that is computationally expensive. Therefore, in order to find quality anchors, we define its goodness on the hypothesis that, "a good anchor pair should align well with many other good anchor pairs". This idea is inspired by the theory of PageRank. To implement this, a graph based scenario is considered where a pair is referred to as vertex (V_i') and the degree of alignment between them defines their weight. The formula to calculate the alignment between pairs is shown in Equation 8.3.

$$\psi_{ij} = cos((\vec{a}_i - \vec{a}_j), (\vec{c}_i - \vec{c}_j)) \tag{8.3}$$

where ψ_{ij} denotes the two pairs (a_i, c_i) and (a_j, c_j). Here, (a_i, a_j) $\in S_a$ (i.e., concepts similar to A) and (c_i, c_j) $\in S_c$ (i.e., concepts similar to C). The intuition behind this is that the difference in vector points of concepts captures the relational/functional alignment between concepts and it is important to preserve this geometric arrangement to precisely capture the query specific semantics.

Equation 8.4 is used to calculate the final weight of each pair. Specifically, the importance (λ) of each pair in the candidate set is computed in a way similar to TextRank algorithm [202] by interactively computing Equation 8.4 until convergence. One crucial advantage of using the idea PageRank is that it promotes pairs with higher authority; as a result, those pairs that have higher connectivity are assigned higher weights. Commonly, generic pairs tend to have higher connectivity than specific pairs. This ensures the pairs that are generic (correspondingly having relatively stable semantic meaning) and simultaneously cognizant to the semantics of input query have a higher impact on the transformation matrix being learned. Algorithm 4 provides the pseudo-code for generating anchor pairs.

$$\lambda(V_i') = (1-d) + d \sum_{V_j' \in Neigh(V_i')} \frac{\psi_{ji}}{\sum_{V_k' \in Neigh(V_j')} \psi_{jk}} * \lambda(V_j') \tag{8.4}$$

where $Neigh(V_j')$ denotes the neighbours of V_j' and d is the damping factor set to 0.85 by default.

Query biased transformation

Based on Section 8.4.3 and Section 8.4.3, we have identified a set (Q) of quality anchor pairs. Now given that, this section enumerates the process to learn the transformation that is sensitive to the relationship between input query terms. Towards this end, the model builds upon some of the special features provided by word embedding spaces such as linear analogical reasoning *vec("ibuprofen") - vec("pain")* $\approx$ *vec("treats")*. In

particular, to capture the relation between anchor pair (a, c), where a is a term similar to 'A' and c is a term similar to 'C', we take the difference of their vector representations. Such linear operations are expected to capture the relational/functional aspect of input query. Our intuition behind this is to preserve the geometric arrangements of pairs in vector space that in turn is expected to encapsulate the precise temporal dynamics particular to a given query. The optimization function for learning local transformation M_2 is given in Equation 8.5.

$$M_2 =_{M_2} \left(\sum_{i=1}^{Q} ||M_2.\lambda_i^0(\vec{a}_i^0 - \vec{c}_i^0) - \lambda_i^1(\vec{a}_i^1 - \vec{c}_i^1)||_2^2 + \gamma||M_2||_2^2 \right) \tag{8.5}$$

where $\vec{a}_i$ and $\vec{c}_i$ refers to the vector position of a_i and c_i at their respective time-slots. λ_i^0 and λ_i^1 are the weights associated with anchor pairs at t_0 and t_1 respectively. The λ_i in Equation 8.5 is the weight associated to each anchor pair based on its "goodness" as compared to other pairs (Using Equation 8.4). Similar to global transformation, the value of regularizer component (γ) is set to 0.02. By default, all the anchor pairs generated are chosen as the size of Q.

Combining with global transformation

Our contention is that the temporal change captured by both global and local transformation has valuable information and their amalgamation is necessary to find important bridge concepts. While the global transformation effectively captures the general information present in the corpus, it misses the subtle cues from the local context. On the other hand, relying solely on query specific transformation risks awarding undue importance to overly specific terms. Thus, it is important to leverage the benefits provided by two distinct but complementary transformations. Against this backdrop, we propose to combine Equation 8.2 and Equation 8.5 and jointly minimizes the following objective function. This allows us to preserve both the global and local proximity of input query simultaneously.

$$M = \alpha M_1 + (1 - \alpha)M_2 \tag{8.6}$$

It can be observed that the final expression still results in regularized least square form. Thus, similar to solving Equation 8.2, we find its closed form updates and obtain the unified transformation matrix. Despite its simplicity, this concatenated approach of linear transformation method worked well in our experiments. The value of α is set to 0.5 by default.

Algorithm 4 GENERATE CANDIDATE ANCHOR PAIRS

Input: Set of concepts similar to A - $setSimilarConcepts(A)$ and Set of concepts similar to C - $setSimilarConcepts(C)$ (From Algorithm 3)
Output: A ranked set (Max Heap of Q) of pair of terms $\{a,c\}$ which are similar to A and C - $candidateAnchors$
$candidateAnchors \leftarrow \phi$
$\{S_a\} \leftarrow setSimilarConcepts(A)$
$\{S_c\} \leftarrow setSimilarConcepts(C)$
$filteredCandidateAnchors \leftarrow \phi$
for $a \in S_a$ **do**
    for $c \in S_c$ **do**
        if ($cosine(a,c) \approx cosine(A,C)$) **then**
            $filteredCandidateAnchors \cup \{a, c\}$
        end if
    end for
end for
$tempCandidateAnchors \leftarrow \phi$
for $pair1 \in filteredCandidateAnchors$ **do**
    for $pair2 \in filteredCandidateAnchors$ **do**
        $align$=$calculateAlign(pair1, pair2)$ //According to equation 8.3
        $tempCandidateAnchors \cup \{pair1\}$
    end for
end for
$candidateAnchors$=$pageRankScore(tempCandidateAnchors)$ // According to equation 8.4
Return $candidateAnchors$

8.4.4 Scoring Conceptual Bridges

Given two previously disconnected terms A and C along with a cut-off time-stamp t' (a meta-constraint to restrict the search space), the goal is to identify plausible bridge concepts k that will connect them in future (t'+1). The candidate for B terms are all the concepts present in vocabulary besides - A, $setSimilarConcepts(A)$, C and $setSimilarConcepts(C)$. Recall that our objective to find bridges concepts that are not

only connected to the query pairs but also to their semantically similar local neighbours. To compute the semantic relatedness of bridges, we first learn the transformation matrix (M) particular to this input query between an initial time stamp t'_0 (by default set to t'-10) and t'. This matrix is learned by the methods described in Section 8.4.3 and is expected to encode the temporal dynamics. Note that as the goal is to predict which conceptual bridge has the highest likelihood at $t'+1$, the corresponding embeddings ($\vec{b}^{t'+1}$, $\vec{a}^{t'+1}$ and $\vec{c}^{t'+1}$) are not available. The following formula is used to compute the likelihood score for each candidate bridge concept (b_k).

$$Score(b_k^{t'+1}) = \frac{1}{2}\{\sum_{i=1}^{N_1} cos(M.\vec{a}_i^{t'}, M.\vec{b}_k^{t'}) + \sum_{j=1}^{N_2} cos(M.\vec{c}_j^{t'}, M.\vec{b}_k^{t'})\} \tag{8.7}$$

where N_1 and N_2 refers to the number of neighbours of A and C, $a_i \in setSimilarConcepts(A)$ and $c_j \in setSimilarConcepts(C)$. Based on the obtained likelihood score, the candidate bridge concepts are ranked and presented to the user.

8.5 Experiments

The focus of this section is to demonstrate the efficacy of the proposed model through a variety of experiments performed under different settings. In our experiments, we use MEDLINE[2] as our main corpora because it provides access to more than 100 years of time-stamped scientific articles, primarily, from life sciences and bio-medicine. The latest dump (2017) contains more than 24 million articles. Every article contains a unique identifier (PMID), title, abstract, publication date and Medical Subject Headings (MeSH) terms. As a unit of representation for articles, we choose MeSH terms. MeSH terms are the special keywords assigned by subject matter experts to each article in MEDLINE. Since these terms are selected by subject matter experts based on the full text of articles, it is safe to assume that they represent the conceptual meaning of an article without adding noise [9, 198]

DataSets: To evaluate the performance of proposed model and compare them with existing hypotheses generation algorithms, the following test cases are chosen. These test case are widely regarded as the "golden dataset" in this area of study [9, 198, 196, 10, 11]. The test cases are enumerated below:

1. Fish-oil (FO) and Raynaud's Disease (RD) (1985)

[2] The source code of Concepts-Bridges is available at `https://github.com/kishlayjha/Concepts-Bridges`.

2. Magnesium (MG) and Migraine Disorder (MIG) (1988)
3. Somatomedin C (IGF1) and Arginine (ARG) (1994)
4. Alzheimer Disease (AD) Indomethacin (INN) (1989)
5. Schizophrenia (SZ) and Calcium - Independent Phospholipase A2 (PA2) (1997)

For the consideration of being self-contained, we briefly provide a background about these test cases. The pioneers in this area of study [5, 78] applied their hypotheses generation technique and postulated above enumerated hypotheses. Later, these hypotheses were clinically verified in the real world laboratories. Since then the re-discovery of these test cases is widely adopted as a way of demonstrating the effectiveness of proposed approach. Note that the dates given for above test cases acts as a threshold to base our analyses. These are the dates when the association between query terms were known and published in the literature. We run the proposed model and baseline algorithms to generate possible connections using all the articles before threshold (pre-cutoff) and then check their validity in the articles present in post-cutoff period.

Evaluation scheme: We provide both qualitative as well as quantitative validation of our approach. In qualitative evaluation, we present the top-k bridge terms and inspect their correctness. In quantitative evaluation, we compare our approach against a variety of baselines and show the superiority of our approach.

Evaluation baselines for quantitative evaluation: To evaluate effectiveness of the proposed model, the following five previous hypotheses generation algorithms are implemented. The initial four algorithms are based on raw term co-occurrence frequency and fifth is a word embedding based approach.

1. *Apriori algorithm*: This algorithm [196] uses two important measure of association rule: a) support and b) confidence to rank the bridge concepts. The threshold for support and confidence are chosen as suggested in [196].

2. *Chi Square ($\tilde{\chi}^2$)*: This study [203] uses Chi-square test to quantify and rank the bridge terms. The threshold for $\tilde{\chi}^2$ is used as suggested in [203].

3. *Term-frequency and Inverse-document frequency (TF−IDF)*: TF-IDF is a popular metric that measures the importance of a concept present in an article. [198] adopts this measure to identify the bridge concepts.

4. *Literature Cohesiveness (coh)*: Literature Cohesiveness is a metric proposed by [204], to identify bridge concepts based on the cohesion of literature.

5. *Static embeddings (Static)*: This algorithm [205] generates cumulative year-wise co-occurence matrix and applies SVD to generate word embeddings. Based on these embeddings, the bridge terms are then ranked using cosine measure.

It should be noted that a direct comparison with the results of above enumerated baselines cannot be performed. This is because of the difference in choice of input, threshold used to select linking terms and the use of domain expertise to prepare gold standard. Nevertheless, to facilitate a fair comparison, their methods have been adjusted to fit the current problem setting.

Evaluation metrics for quantitative evaluation: Two evaluation metrics are used to quantify our results: 1) Precision@k and 2) Mean Average Precision (MAP). Precision@k allows us to measure the coverage of ground truth terms in top-k target set; thus allowing analysis at a granular level. To quantify the system's performance across queries, we report MAP.

8.5.1 Qualitative evaluation

In this section, we evaluate our proposed model based on its ability to rediscover the existing knowledge.

Fish-Oil - Raynaud's Disease: In this test case, the pioneers identified that fish oils might prevent raynaud disease by a) inhibiting platelet aggregation, b) reducing blood viscosity and c) preventing vasoconstriction (epoprostenol) [5] and reported them in an article in 1986. These *conceptual bridges* were later experimentally validated. In our experiments, we seed our algorithm with input pairs (A,C) as ("fish oils", "Raynaud disease") and a date (t') as 1985. T

Migraine - Magnesium: The objective of this test case was to examine the effect of magnesium in treating migraine disorder. Similar to the previous case, several intermediate terms such as *epilepsy, serotonin, prostaglandins, platelet aggregation, calcium antagonist, type A personality, vascular tone and reactivity, calcium channel blockers, spreading cortical depression and substance P* were reported. Unlike the previous case, we are unable to achieve high recall. Nevertheless, we obtain important conceptual bridges such as *epilepsy*, *calcium antagonist*, *prostaglandins*, etc. Note that the previous studies indicate this to be a difficult test case [198].

Indomethacin - Alzheimer Disease: The most significant pathways reported for this case are *Acetylcholine* and *Membrane fluidity*. Both of these pathways were found

in top five.

Somatomedin C - Arginine: For this test case, *Somatotropin* and *somatostain* are the most important pathways [204]. In our results, we were able to obtain both of them in top five.

Schizophrenia - CI Phospholipase A2: The initial studies reported oxidative stress to be the key connecting term for this test case. In our results, we found Dopamine Receptors (a derivative of oxidative stress at rank 3).

Overall, the proposed model was able to identify a majority of true connections at top ranks, however, a related questions arises: How novel are the other top terms reported?

Discovery Example: For the first test- case (FO-RD), one of the term reported in Top-10 was *beta-thromboglobulin*. Beta-thrombo-globulin is a platelet-specific protein that is released when platelets aggregate. Manually inspecting the literature, we found that an article [206] in 2001 reported the potential role of beta-thromboglobulin in preventing endothelial cell damage that is known to cause Raynaud's disease. Although prior to 1986 there was no reported connection, the proposed model could identify it by analyzing existing connections in the medical literature. Similarly, for another test-case of INN-AD, one of the top ranked connecting term was *Phenylacetates*. More recently, [207] reported the potential role of Phenylacetates in treatment for Alzheimer Disease. While these connections are being reported recently in the literature, the model was able to identify them much in advance. We believe one reason for this lies in the choice of modelling in latent space that enables the algorithm to find connections that might be surprising at the time of being postulated. To further aid the biomedical scientists in conducting extensive study, we provide evidence for our top 10 terms (refer Table **??**) in the form of PMID.

Based on the rediscovery of existing knowledge and aforementioned discovery scenario, it can be deduced that the model is able to replicate already known knowledge and possibly originate new knowledge. However, this form of evidence based evaluation does not inform us about the overall quality of result set. To this end, a quantitative evaluation has to be performed.

Table 8.1: Precision@k for FO-RD

Algorithm	k=10	k=20	k=30	k=40	k=50
apriori	0.5	0.6	0.6	0.55	0.54
TF-IDF	0.4	0.6	0.567	0.55	0.54
$\tilde{\chi}^2$	0.4	0.4	0.567	0.475	0.54
coh	0.6	0.5	0.467	0.5	0.5
static	0.2	0.35	0.467	0.45	0.46
Concepts-Bridges	**0.8**	**0.7**	**0.667**	**0.625**	**0.62**

8.5.2 Quantitative evaluation

The purpose of this section is to probe the overall quality of output generated. However, to perform a quantitative analysis certain ground truth is required. Unfortunately, there is no standard ground truth available and creating one remain an open problem [208]. One reason behind this is the fact that it is near-impossible to build a comprehensive ground truth set that will presumably have all the future discoveries. Therefore, a "supposedly" ground truth has to be constructed. To accomplish this goal, a split corpus approach is adopted. Specifically, the dataset is divided into two sets: 1) Pre-cut-off segment: this includes articles published before the cut-off date and 2) Post-cut-off segment: this includes articles published after the cut-off date. The proposed model and baseline algorithms are run on the pre-cut-off segment. Then, the generated connections are checked in the post-cut-off segment. The legitimacy of a connection is defined as its presence (co-occurence) in post-cut-off segment and absence in pre-cut-off. Equation 8.8 presents the formula to rank ground truth bridge term k for a given pair (A, C).

$$gt(k) = \frac{\#(k, A) + \#(k, C)}{\#(k)}, \tag{8.8}$$

where $\#(i, j)$ is the number of times terms i and j co-occur and $\#(i) = \sum_j \#(i, j)$. In this way, a ranked set of ground truth is constructed. As a post-processing step, all the stop-words (also referred to as check-tags in medical domain) are removed from the resultant set.

Results: Table 9.1, 9.2, 9.4, 9.3, 9.5 reports the Precision@k for each of the five golden datasets. The value of K is gradually increased from 10 to 50 (in the interval of 10) and results are reported. Table 9.6 reports the Mean Average Precision @k by

Table 8.2: Precision@k for MG-MIG

Algorithm	k=10	k=20	k=30	k=40	k=50
apriori	0.7	0.65	0.7	0.675	0.64
TF-IDF	0.8	0.65	0.7	0.66	0.66
$\tilde{\chi}^2$	0.4	0.55	0.667	0.6	0.62
coh	0.5	0.45	0.5	0.525	0.54
static	0.5	0.55	0.633	0.675	0.66
Concepts-Bridges	**0.8**	**0.8**	**0.733**	**0.725**	**0.7**

Table 8.3: Precision@k for AD-INN

Algorithm	k=10	k=20	k=30	k=40	k=50
apriori	0.6	0.7	0.8	0.75	0.66
TF-IDF	0.5	0.55	0.7	0.75	0.7
$\tilde{\chi}^2$	0.6	0.65	0.667	0.675	0.64
coh	0.6	0.7	0.7	0.7	0.66
static	0.7	0.65	0.7	0.675	0.7
Concepts-Bridges	**0.9**	**0.85**	**0.833**	**0.825**	**0.8**

Table 8.4: Precision@k for IGF1-ARG

Algorithm	k=10	k=20	k=30	k=40	k=50
apriori	0.8	0.85	0.833	0.725	0.7
TF-IDF	0.5	0.45	0.467	0.575	0.64
$\tilde{\chi}^2$	0.6	0.7	0.7	0.7	0.7
coh	0.8	0.85	0.833	0.825	0.7
static	0.6	0.45	0.433	0.525	0.58
Concepts-Bridges	**0.9**	**0.9**	**0.867**	**0.85**	**0.84**

Table 8.5: Precision@k for SZ-PA2

Algorithm	k=10	k=20	k=30	k=40	k=50
apriori	0.6	0.75	0.767	0.825	0.82
TF-IDF	0.4	0.6	0.7	0.75	0.78
$\tilde{\chi}^2$	0.5	0.7	0.767	0.825	0.86
coh	**1.0**	0.95	**0.967**	0.85	0.82
static	0.4	0.6	0.7	0.775	0.78
Concepts-Bridges	**1.0**	**1.0**	**0.967**	**0.95**	**0.92**

consolidating numbers across different datasets.

Discussion: It can be observed that the proposed model outperforms all the existing baselines. Across all the datasets, a common pattern noticed for the proposed model is the decrease in precision with the increase in value of K. In contrast, for baseline algorithms the precision increases (in general) with increase in value of K. This trend elucidates the advantage of proposed model to rank relevant connections at higher positions. Analyzing the results further, we observe that Literature Cohesivenes (COH) performs the best among all the baselines. Perhaps, the reason for this lies in the ability of COH to leverage the cohesion of literature effectively.

Another important point to note is that pure frequency based approaches (Top 4 baselines) boosts contextually generic terms at higher positions. Contextually generic terms are those terms that are generic to the given input query. For instance, in the Fish Oils and Raynaud's disease test case, some of the top terms found for COH (and other baselines) are "double-blind method", "skin ulcer", "leg ulcer" and so on. Although these terms have relatively lower overall frequency they tend to frequently co-occur

Table 8.6: Mean Average Precision@k for all test cases

Algorithm	k=10	k=20	k=30	k=40	k=50
apriori	0.616	0.667	0.702	0.728	0.744
TF-IDF	0.538	0.571	0.594	0.615	0.632
$\tilde{\chi}^2$	0.477	0.548	0.587	0.605	0.618
coh	0.731	0.723	0.725	0.723	0.725
static	0.442	0.487	0.528	0.552	0.570
Concepts-Bridges	**0.907**	**0.907**	**0.860**	**0.847**	**0.836**

with the input query (i.e., Raynaud's disease). Selecting these terms prove counter-productive as they are ranked lower in the ground truth. The reason being, these contextually generic terms have no functional relationship with the input concept. Note that more often the true conceptual bridges have important functional relationship with input query. For example: Fish oils $\xrightarrow{\text{disrupts}}$ platelet aggregation $\xrightarrow{\text{cause}}$ Raynaud's disease. Furthermore, as the fifth baseline (Static embeddings) too does not factor in the "functional" aspect, it suffers from this issue. To mitigate these issue, the proposed model (in particular query-specific transformation component) takes advantage of the analogical relationships provided by word embedding spaces to capture the functional component of medical concepts.

Another reason for the lower performance of baselines lies in the fact that they strictly rely on the given input query (ignoring the cues from local neighbourhood). This is limiting because more often when a potential conceptual bridge (e.g. "platelet aggregation") is being studied/reported in the literature with a particular concept (e.g. "fish oils"), it is highly likely that it is also being reported with the chemical substances/genes associated with them. In this case, the chemical substance being "eicosapentaenoic acid". Models based on strict query reliance attempt to find bridge concepts ("platelet aggregation") by only considering the semantic association with particular input concept ("fish oils"). Ignoring such semantically similar neighbours ("eicosapentaenoic acid") may limit the capability of model to find potential bridge concepts. Note that some of the existing approach [10] manually augment their input query to enrich their relevant document set. However, this requires the user to possess some form of domain knowledge. In the proposed approach, the use of word embeddings automatically enables to find semantically similar concepts that enriches the user provided input queries. Lastly,

Table 8.7: Effect of global and local transformation. MAP@K

Algorithm	k=10	k=20	k=30	k=40	k=50
global	0.728	0.715	0.697	0.688	0.657
local	0.816	0.797	0.782	0.781	0.764
Concepts-Bridges	**0.907**	**0.907**	**0.860**	**0.847**	**0.836**

the fifth baseline chosen for comparison is Static embeddings (Static). This baseline ranks the bridge concepts based on the static embeddings generated from cumulative co-occurrence matrix. Our intend behind this is to test the necessity of leveraging temporal dynamics itself. Static essentially assumes a static world in which each term is supposed to retain its semantics across different domains. As reported in the results, we can see that the proposed approach outperforms it. This result suggests that it is crucial to consider the temporal change of concepts in order to generate semantically sensible hypotheses.

8.5.3 Effect of global and local transformation

The only parameter in the proposed approach is the α in Equation 8.6. The α parameter controls the contribution of global and local transformation. Table 8.7 compares the influence of each transformation in the form of MAP@k calculated for all the five test cases. As can be seen, the local transformation outperform global transformation. We believe the reason for this lies in the ability of local transformation to encode query-specifics semantics in an effective manner. Furthermore, the best result comes from combination of both global and local, thus validating the need for Equation 8.6.

8.6 Conclusions

In this study, we proposed a new model to discover conceptual bridges between two disparate but complementary topics of inquiry. Specifically, the model leverages upon the temporal information present in the corpus and captures the semantic change of medical concepts at a coarse-grained level. The proposed query-biased transformation technique, in particular, leverages the fundamental relationship between input query and its informative neighbours to encapsulate precise semantics. This enables the model to promote those conceptual bridges that have higher semantic meaning. Empirically,

we evaluate the model in a variety of experimental settings. The experimental results demonstrate that the proposed model has the potential of generating practical new knowledge. In future research, we intend to add more semantic expressiveness to our generated hypotheses. Towards this end, we are looking at more specialized biomedical resources such as SEMMEDDB - a repository of semantic predications in the form of 'subject-predicate-object'.

Chapter 9

Hypothesis Generation based on Co-Evolution of Biomedical Concepts

9.1 Introduction

The constant influx of scientific articles and their easy accessibility via the World Wide Web (WWW) has made medical informatics a fast growing field [3]. Practitioners in the field have thrived to make sense of huge number of academic publications, discovery notes, electronic medical records and other text materials (a.k.a "big biomedical data") leading to advancements of practical significance [4]. While this swift availability of scientific information has acted as an impetus for pacing research innovation, at the same time, it has also overwhelmed researchers trying to survey published studies and construct novel research hypotheses. For instance, consider a novice researcher attempting to formulate a new hypothesis for the cures of *Diabetes*. In doing so, at this point in time, one might have to survey tens of thousands of existing publications (more than 400,000 in PubMed [3] alone) already written on *Diabetes*. This overloaded amount of information presents a bottleneck, as it is almost impossible for one to process and analyze such a large volume of available material. Moreover, it introduces delays in scientific productivity, as biomedical researchers are faced with a daunting task of choosing postulates/hypotheses - based upon the manual inspection of literature - for possible

in-vitro clinical trials. To mitigate these issues, there has been a growing research interest among data/text mining practitioners to develop computational models that are able to assist biomedical experts in forging analytically probable and medically sensible hypothesis. Towards this end, Hypotheses generation (HG), a sub-problem of biomedical text-mining, aims to discover cross-silo connections (also known as undiscovered public knowledge) by chaining together the already known and established scientific facts that remain dispersed across the disparate research fields [8]. Simply put, given an input concept of interest (e.g., disease or gene), HG attempts to find implicit links (e.g., potential drug target or novel indicator of disease's mechanism) that connects them in a previously unknown but semantically meaningful way. Finding such meaningful associations is the crux of the problem that this paper attempts to address.

Over the past few decades, numerous studies have been conducted to tackle this problem. Broadly, they can be categorized into three major groups: a) distributional approaches [6, 9], b) graph-based methods [10, 11] , and c) supervised machine learning based approaches [12]. Arguably, these studies made significant advances, however, they still contain a few inherent drawbacks. First, a majority of these preceding approaches rely on a pre-defined structure (e.g., graph) and hence possibly risk missing links that are not included in their route. Second, almost all of these studies assume that the domain is static. This is limiting because it is known that the biomedical domain is a highly evolving field with new facts being added and old ones being obsolete every single day [13]. To overcome these issues, more recently, a few studies [7, 23] attempted to formulate this problem in latent space and generated hypotheses by modeling the temporal evolution of concepts based on the diachronic biomedical corpora. While these studies substantiated the importance of leveraging the temporal component, they still neglected the evolutionary features of concepts present in contemporary biomedical ontologies. Such ontologies/taxonomies in biomedical domain are constantly updated by subject-matter-experts to reflect the *up-to-date* knowledge of the field. Thus, to gain a holistic understanding of temporal change, it is crucial to factor in the semantic change of medical concepts from these subject-matter-experts maintained KB too. Furthermore, in practice, a significant amount of information is also encoded in the (co)-evolutionary dynamics of medical concepts between these complementary sources of information (i.e., corpus and ontology). Considering the complementary strength of both these resources, a few natural questions arise: Would the joint modelling of co-evolutionary dynamics lead to the generation of robust temporal embeddings? Would the mutual interaction

between these intertwined resources simulate better predictive effects and thus benefit tasks such as hypothesis generation? To answer these questions, in this study, we model the co-evolution of medical concepts driven by the complex interaction between concepts' linguistic usage (reflected in local context information) and their structural localities (reflected in domain ontology). More specifically, we achieve this by adopting a shared temporal matrix factorization framework, wherein the subspaces between multiple related matrices are jointly learned by sharing information between them. By collaboratively exploiting the evolutionary features of medical concepts from both corpus and domain knowledge, the proposed approach yields hypotheses that are medically sensible and of potential interest to the domain experts. In this study, our contributions can be summarized as:

1. We propose a general framework for the task of hypothesis generation that is capable of inferring previously unknown but potentially interesting cross-silo connections by capturing the subtle cues manifested in the temporal drift.
2. The proposed approach for capturing the temporal change models the (co)-evolutionary dynamics of medical concepts across both the complementary sources of information - corpus and domain knowledge - thereby generating temporal embeddings that are robust and useful for a variety of downstream biomedical text-mining tasks.
3. We propose an effective technique to leverage the evolving topological properties of biomedical KB, resulting in vector representations that encode the temporal dynamics at a granular level.

9.2 Related Work

Discovering hidden, previously unknown and potentially useful associations between biomedical concepts is a problem of practical value in the research area of biomedical text-mining [209, 164, 210, 211, 212, 164, 134, 213, 214, 65, 215, 216, 217, 218, 219, 220, 135, 136, 115, 23, 177, 221, 222, 223]. For a recent survey, please refer [221, 223]. The initial works [5, 8] in this area of study elucidated that the novel implicit links (e.g., *Fish Oils* $\xrightarrow{\text{treats}}$ *Raynaud's disease*) can be discovered by connecting independent nuggets of information remaining dispersed across the literature. While these pioneering studies laid the foundational groundwork, they were extremely time-consuming. Consequently, the subsequent studies focused on automating it. Primary studies such as [6, 9] applied

statistical co-occurrence techniques (term frequency, inverse document frequency, record frequency and so on) to quantify the statistical strength between links. Similarly, [196, 9] adopted associate rule mining technique to estimate the strength of co-occurrences between concepts. While these purely co-occurrence based methods were progressive, a major drawback lies in their over-reliance on term frequencies. A greater statistical association implies strong but not necessarily semantically meaningful (real biological significance) association. To circumvent this drawback, we choose to model the problem of HG in latent space wherein the system is capable of capturing the implicit semantics between concepts, thereby finding connections that have greater semantic association.

Meanwhile, another line of research focused on modeling the problem of HG in a graph-based setting. Since graph based methods provide a natural way of representing concepts and their relationship, this line of research has attracted considerable attention. In [11], the authors presented a novel graph-based approach utilizing semantic predicates (subject-predicate-object), where subject/object refer to the entities (nodes) and predicates refer to the relationship (edge) between them. Another popular graph based HG system is *Obvio* [10]. Given a user input, *Obvio*, first constructs a graph on-the-fly and then uses the context information to automatically create semantically meaningful sub-graphs. One major contribution of this study is their ability to elucidate the meaning of complex associations between medical concepts along the multiple thematic dimensions. While graph-based approaches [11, 10] remain more successful than their distributional counterparts, they are still unable to find implicit connections. This is because the graph-based techniques still rely on a pre-defined structure/schema. More recently, some of the studies such as [12] applied supervised machine-learning based techniques to find the hidden connections. However, they require the domain expertise to generate the training data. This is both time-consuming and monetarily expensive. Despite important advances made, all of the aforementioned studies assumed the biomedical domain to be static. This is limiting because the domains in general (and in particular biomedicine) are usually dynamic with updates being made every now and then. To overcome this issue, recently, a few studies [23, 7] incorporated the temporal component by modelling the semantic evolution of medical concepts present in the historical biomedical corpus. However, these studies still neglect the semantic change of concepts from KB and thus fail to leverage the (co)-evolutionary dynamics of medical concepts.

Some of the motivation for this study stems from the research area of temporal

network modelling [224].While close in spirit, we differ from them in two aspects: a) Our focuses are different. b) Unlike modelling the temporal dynamics from multiple views of a network, in the current problem setting, our objective is to model the (co)-evolutionary features of medical concepts from their linguistic usage and structural localities in a concurrent manner.

9.3 Methodology

In this section, we describe our proposed framework in detail. Recall that the input to our hypothesis generation system is a topic of interest (A), date (d), and the goal is to predict previously unknown implicit links (C) at (d + 1). To tackle this problem, the key intuition behind our proposed approach is the following: If two medical concepts (A and C) are known to be primarily disjoint (i.e., no known relationship exists), and yet their implicit semantics continue to grow closer to each other over time, then these two terms have a higher chance of materializing a meaningful connection in the near future. In other words, our core objective is to capture the temporal 'proximity' between the medical concepts by modelling their semantic change over time. Generally speaking, this can be achieved by adopting a two-step solution: a) apply the temporal word embedding model [23] and generate the time-aware vector representations of concepts, b) quantify the degree of proximity between concepts by measuring the distance between their vector representations. While effective in practice, this class of techniques generate temporal embeddings in an isolated manner (e.g., corpus/ontology alone), and thus neglect the prevalent (co)-evolutionary features of medical concepts. To overcome this, in this study, we aim to generate the temporal embeddings that are infused with (co)-evolutionary dynamics generated due to the mutual influence of both complementary sources of information - corpus and ontology. Technically, we achieve this by adopting a shared temporal matrix factorization framework, wherein the sub-spaces between multiple related matrices are mutually learned by sharing the information between them. Further details on this are provided in the subsequent sub-sections. Section 9.3.1 and Section 9.3.2 introduce the two building blocks (modelling corpus-based and ontology-based evolutionary dynamics) of the proposed model. Then, in Section 9.3.3, we describe the joint co-evolution framework.

9.3.1 Corpus-Based Evolutionary Dynamics

To obtain the corpus-based temporal embeddings, we first need a text corpus collected across time (e.g., time-stamped scientific articles). Given this corpus, the objective is to generate the temporal word embeddings for each word present in the corpus. Traditionally, these temporal word embeddings could be generated by applying the neural network inspired language models such as Skip-gram (augmented with temporal component) [16] to the input sequential text. Simply put, the objective function of skip-gram is to predict the surrounding words within a fixed window, given a focus word. Following similar research direction, more recently in a related study [77], the authors proved that the objective function that the neural network attempts to solve in case of Skip-gram model with negative sampling is the same as the matrix factorization of Shifted Positive Point-wise Mutual Information (SPPMI) matrix obtained from the co-occurrence matrix of the corpus. As a result, the word and its corresponding context vectors can be obtained from the matrix decomposition of SPPMI matrix. This result is attractive as it enables the adoption/extension of techniques from the well-established area of matrix factorization. In this study, we utilize this equivalence result and propose a temporal matrix factorization based framework to obtain our temporal embeddings.

Formally, let us denote D_t as our time-stamped text corpus, where time-stamp t represents a discrete and ordered variable that varies from 1 to T. Given this corpus, we first collect all the concepts occurring in the corpus and prepare an overall vocabulary $V = \{w_1, ..., w_v\}$ of size $|V|$, where each w_i corresponds to an individual term. Note that this vocabulary is common to both the corpus and chosen ontology. Next, we construct a term-by-term $\mathbf{Y}(t)$ Pairwise Mutual-Information Matrix (PMI), whose i, j-th entry is:

$$PMI(i,j)_t = \log\left(\frac{\#(i,j)_t.|D_t|}{\#(i)_t.\#(j)_t}\right) \tag{9.1}$$

where $\#(i,j)_t$ counts the number of times the words w_i and w_j co-occurs within a document over the corpus D at time t, $\#(i)_t$ and $\#(j)_t$ denotes the total number of times words w_i and w_j occur in the corpus at time t alone. $|D_t|$ is the total number of word tokens in the corpus at time t. Following this, we compute the shifted positive point-wise mutual information matrix (SPPMI) specific to a corpus D at time t, whose (i,j)-th entry is:

$$SPPMI(i,j)_t = max(PMI(i,j)_t - \log k, 0) \tag{9.2}$$

where $\log k$ refers to a global constant. The constant $\log k$ acts as a prior on the probability of observing a positive example versus a negative example. A higher value of k implies that negative examples are more likely.

Following this idea, now our objective is to obtain a dense, low-dimensional vector representation $\mathbf{V}^{'}(t) = [\mathbf{v}'_{w_1}(t), \mathbf{v}'_{w_2}(t), ..., \mathbf{v}'_{w_v}(t)] \in R^{|V| \times n}$, $n \ll |V|$ for each word $w \in V$, at each time-period t. $\mathbf{v}'_{w_i}(t)$ denotes the embedding vector for the i-th word at time-stamp t, and n is the number of dimensions. To achieve this, we adopt a standard matrix factorization framework and set up a least square optimization problem, so that the PPMI matrix $\mathbf{Y}(t)$ matches $\mathbf{U}.\mathbf{V}^{'}(t)^T$ as closely as possible. The formulated optimization is shown below:

$$\min_{\mathbf{U}, \mathbf{V}'(t) \geq 0} \sum_{t=1}^{T} \frac{h(t)}{2} ||\mathbf{Y}(t) - \mathbf{U} \cdot \mathbf{V}^{'}(t)^T||_F^2 \tag{9.3}$$

Both $\mathbf{U}$ and $\mathbf{V}'(t)$ are $|V| \times n$ matrices. The main difference between $\mathbf{U}$ and $\mathbf{V}'(t)$ is that $\mathbf{U}$ is a constant matrix and $\mathbf{V}'(t)$ is a time-dependent matrix. While it is possible to make both $\mathbf{U}$ and $\mathbf{V}'(t)$ time-dependent, as shown in [224], a simpler model can achieve good approximation and also avoid over-fitting. The function $\mathbf{V}'(t)$ can take on any canonical form, such as linear models, polynomial models and so on. $h(t)$ refers to a decay function that regulates the importance between current and historical snapshots. This acts as an smoothing. The exponential function is chosen as a decay function with parameter $\theta > 0$.

$$h(t) = e^{-\theta(T-t)} \tag{9.4}$$

One challenge in this setting is that the PPMI matrix $\mathbf{Y}(t)$ is large and difficult to fit into memory. However, as most of the real-world networks are usually sparse, the computation can be made efficient. In most of the real-world scenarios, the presence of a co-occurrence conveys more significant information than the absence of a co-occurrence. This is because the absence of a co-occurrence could mean: a) either there exists no association between the two concepts. b) there might exist a possible association between them in the near future. The presence of co-occurrence is seemingly more meaningful and thus the aforementioned objective function is adjusted to prioritize the presence of co-occurrence rather than the absence of co-occurrence. However, a small number of negative co-occurrence is needed to properly train the model. Suppose $E(t)$ be the set of word-pairs (w_i, w_j) such that the value of $y_{ijt} = 0$, and $F(t)$ be the set of word-pairs

(w_i, w_j) such that the value of $y_{ijt} > 0$. Then, total set of co-occurrences is shown below:

$$G(t) = E(t) \cup F(t) \tag{9.5}$$

Now, one can express the objective function as:

$$\min_{\mathbf{U},\mathbf{V}'(t)\geq 0} \sum_{t=1}^{T} \frac{h(t)}{2} \sum_{(w_i,w_j)\in G(t)} (y_{ijt} - (\mathbf{U} \cdot \mathbf{V}'(t)^T)_{ij})^2 \tag{9.6}$$

Note that non-negativity is imposed on the factors for the purpose of greater interpretability.

9.3.2 Ontology-Based Evolutionary Dynamics

Ontologies/Hierarchies usually represented as Trees are known to provide a natural way of categorizing the knowledge of a particular domain. Such ontologies, also referred to as knowledge-bases (KB's), are abundantly present in the biomedical domain. Some common examples include Medical Subject Headings (MeSH), Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT), and International Classification of Diseases (ICD9). These KB's are periodically updated by the subject-matter-experts in order to reflect the contemporary knowledge of the field. Given that these KB's are manually curated and showcase the prevailing knowledge of the field, our speculation is that integrating the evolutionary features of concepts from these resources will result in more accurate temporal representation of biomedical concepts. In our present study, the KB chosen is hierarchical (i.e., IS-A relationships) in nature (further details in experiments). Basically, the edges between concepts in the Tree denotes "parent-child" relationship, and the depth of a concept from the root indicates its level of specificity. Note that greater the depth of a concept in the tree the greater is its semantic richness. To leverage this valuable information, we adopt a technique similar to Section 9.3.1, and later extend our objective function. More specifically, we first convert the given hierarchical KB into a semantic distance matrix $\mathbf{M}(t)$[1] , and then approximate the semantic distance matrix by the product of two smaller matrix.

$$\mathbf{M}(t) \approx \mathbf{U}.\mathbf{V}''(t)^T \tag{9.7}$$

[1] Note that the hierarchical KB is released every year and thus evolves over time.

Here both $\mathbf{U}$ and $\mathbf{V}^{''}(t)$ are $|V| \times n$ matrices with $n << |V|$, n denotes the number of dimensions. The semantic distance matrix $\mathbf{M}(t)$ between concepts is calculated based on two factors: a) shortest path between concepts, and b) the depth of least common subsumer (LCS). The LCS refers to the immediate common parent of two concepts. Given two concepts w_i, w_j at time t, the distance between them is calculated by the formula below:

$$\begin{aligned} l_{ij} = \log_2([\text{path}(w_i, w_j) + 1] \\ * [D' - \text{depth}(lcs(w_i, w_j))]) \end{aligned} \tag{9.8}$$

where $\text{path}(w_i, w_j)$ is the shortest distance between concept w_i, w_j at time t, $\text{depth}(lcs(w_i, w_j))$ is the depth of $lcs(w_i, w_j)$ at time t, D' is the maximum depth of the taxonomy, and $lcs(w_i, w_j)$ is the lowest common subsumer of w_i and w_j. Prior research studies [39] have shown that the exploitation of these two factors is an effective strategy to leverage the ontology specific features. Having obtained our semantic distance matrix $\mathbf{M}(t)$, our next step is to generate the ontology-specific temporal embeddings. To do so, similar to Equation 9.3, the optimization problem is formulated as shown below:

$$\min_{\mathbf{U}, \mathbf{V}''(t) \geq 0} \sum_{t=1}^{T} \frac{h(t)}{2} ||\mathbf{M}(t) - \mathbf{U} \cdot \mathbf{V}^{''}(t)^T||_F^2 \tag{9.9}$$

Though intuitive, in practice, this basic formulation does not fully leverage the typical topological properties of given hierarchical KB. To overcome this issue, we propose an enhanced strategy that exploits the topological properties of the available taxonomy in a more effective manner. Basically, we consider a practical assumption that in the hierarchical KB, *the meaning of a particular concept is particularly influenced by its ancestors in the following order: direct-parents (strongest), grand-parents (stronger), higher-ancestors (lower) and root (least).* As an example, consider the concept "Diabetes Mellitus, Lipoatrophic". This concept forms its semantics by inheriting the basic properties from its ancestor concepts ("Diabetes Mellitus, Type 2", "Diabetes Mellitus", "Endocrine System" and "root")[2] , and also adds its own specific properties. Accordingly, the vector representation of a concept w_i should be modelled by quantifying the semantic contribution for each of its ancestor w_{ij}. We define the strategy to quantify semantic contribution by exploiting the principles of label propagation [225, **?**], usually adopted in network modeling tasks. Simply put, the idea in label propagation

[2] `https://meshb.nlm.nih.gov/record/ui?ui=D003920`

is to preserve the local spatial consistency of network by nudging the neighbourhood concepts to have similar feature vectors. Much alike, we mould its principles to fit the current hierarchical structure of KB, and argue that the features of a concept should be particularly influenced by their ancestors in accordance to their level of specificity.

$$b_{ij}^{(t)} = \frac{1}{\sqrt{\lambda}} \tag{9.10}$$

λ denotes the depth of ancestor concept (w_{ij}) in the tree. Note that the semantic contribution value of each concept changes over time based on their evolving structural locality. Having calculated the semantic contribution value, now, each concept in the tree adjusts (updates) its feature vectors based on its ancestors. Suppose that the initial feature vector of concept w_i is $\mathbf{v}_i(t)$, and the updated vector is $\mathbf{v}^{''}(t)$ at timestamp t. Then, the feature vector update process from $\mathbf{v}_i(t)$ to $\mathbf{v}^{''}(t)$ can be modeled by the following optimization problem.

$$\min_{\mathbf{v}^{''}(t)} \alpha \sum_i ||\mathbf{v}_i^{''}(t) - \mathbf{v}_i(t)||^2 + (1-\alpha) \sum_{j \in Ancestors(w_i)} b_{jj}^{(t)} ||\mathbf{v}_i^{''}(t) - \mathbf{v}_{ij}(t)||^2 \tag{9.11}$$

In the above Equation 9.11, the first term is known as the fitting constraint. This constraint penalizes large deviation from the initial feature vectors. The second term ensures that the feature vectors of concepts are updated in accordance to the semantic contribution of its ancestors. α balances the contribution of each part of the equation. As the formulation in Equation 9.11 is convex, its solution can be found by solving a system of linear equations. The closed updates are give below:

$$\mathbf{v}_i^{''}(t) = (1-\alpha)(\mathbf{I} - \alpha \mathbf{B}(t))^{-1} \mathbf{v}_i(t) \tag{9.12}$$

where $\mathbf{I} \in R^{|V| \times |V|}$ is an identity matrix. $\mathbf{B}(t)$ is defined as the depth matrix. Next, we substitute the analytical solution of Equation 9.11 in Equation 9.9.

$$\mathbf{S}(t) = (1-\alpha)(\mathbf{I} - \alpha \mathbf{B}(t))^{-1} \tag{9.13}$$

$$\mathbf{V}^{''}(t) = \mathbf{S}(t)\mathbf{V}(t) \tag{9.14}$$

$$\min_{\mathbf{U}, \mathbf{V}''(t) \geq 0} \sum_{t=1}^{T} \frac{h(t)}{2} ||\mathbf{M}(t) - \mathbf{U} \cdot \mathbf{S}(t) \cdot \mathbf{V}(t)^T||_F^2 \tag{9.15}$$

In this regard, one might ask: What is the necessity of adopting this route when the semantic distance matrix $\mathbf{M}(t)$ already captures the global hierarchical information? In our research we found two reasons for it: a) the strategy to exploit the typical ancestral property of a given concept acts as a "local regularization" and thus aids to leverage the taxonomic features in a more effective way. b) it provides a good initialization (generates basis vectors that are much closer to the best basis vectors found) for the Non-negative matrix factorization (NMF) formulation, resulting in improved convergence speed and accuracy.

9.3.3 Corpus-Ontology Based (Co)-Evolutionary Dynamics

Both Section 9.3.1 and Section 9.3.2 can obtain the temporal embeddings for biomedical concepts. The former exploits the local context information from natural language text and the later leverages upon the topological properties of given taxonomy. However, these two components should not be isolated from one another as they provide complementary sources of information. Furthermore, a significant amount of information is encoded in their (co)-evolutionary dynamics with respect to one another. To address this, we propose to jointly model the co-evolution of biomedical concepts from these interdependent sources of information. The objective function to be optimized is shown below:

$$\begin{aligned}\min_{\mathbf{U},\mathbf{V}(t),\mathbf{V}'(t)\geq 0} \sum_{t=1}^{T} \frac{h(t)}{2} ||\mathbf{Y}(t) - \mathbf{U}\cdot\mathbf{V}^{'}(t)^{T}||_F^2 \\ +||\mathbf{M}(t) - \mathbf{U}\cdot\mathbf{S}(t)\cdot\mathbf{V}(t)^{T}||_F^2\end{aligned} \tag{9.16}$$

As it can be observed, the first and second part of objective function models the temporal change of concepts from natural language text and ontology respectively. To facilitate the joint learning and mutual sharing of information, the latent factor $\mathbf{U}$ is shared by both parts of the objective function. As mentioned before, both $\mathbf{V}$, $\mathbf{V}'$ can take any canonical form (e.g., linear, polynomial and so on). For simplicity of the model, we choose a linear function. For instance: $\mathbf{V}(t) = \mathbf{X}t + \mathbf{Y}$. As $\mathbf{V}(t) \geq 0$, both $\mathbf{X} \geq 0$ and $\mathbf{Y} \geq 0$. Now, after adding regularization terms the expanded form of Equation 9.16 becomes:

$$J(\mathbf{U},\mathbf{X}',\mathbf{Y}',\mathbf{X},\mathbf{Y}) = \sum_{t=1}^{T}\frac{h(t)}{2}\sum_{(w_i,w_j)\in G(t)}(y_{ijt}-\mathbf{U}\cdot(\mathbf{X}'t+\mathbf{Y}')^T)_{ij}+$$
$$\sum_{(w_i,w_j)\in G(t)}(m_{ijt}-\mathbf{U}\cdot\mathbf{S}(t)\cdot(\mathbf{X}t+\mathbf{Y})^T)_{ij}+\frac{\beta}{2}||\mathbf{U}||^2 \tag{9.17}$$
$$+\frac{\gamma_1}{2}||\mathbf{X}||^2+\frac{\omega_1}{2}||\mathbf{Y}||^2+\frac{\gamma_2}{2}||\mathbf{X}'||^2+\frac{\omega_2}{2}||\mathbf{Y}'||^2$$

where $G(t)$ refers to the set of co-occurrence set as defined in Equation 9.6. The bound-constraint formulation of the above objective function is shown below:

$$\min_{\mathbf{U},\mathbf{X}',\mathbf{Y}',\mathbf{X},\mathbf{Y}} J(\mathbf{U},\mathbf{X}',\mathbf{Y}',\mathbf{X},\mathbf{Y})$$
$$\text{subject to } \mathbf{U},\mathbf{X}',\mathbf{Y}',\mathbf{X},\mathbf{Y}\geq\mathbf{0} \tag{9.18}$$

Next, we find the update rules for our cost function $J(\mathbf{U},\mathbf{X}',\mathbf{Y}',\mathbf{X},\mathbf{Y})$ with respect to each of the model parameters $\{\ \mathbf{U},\mathbf{X}',\mathbf{Y}',\mathbf{X},\mathbf{Y}\ \}$ and run the stochastic gradient descent. The choice of optimization method is agnostic to the model and thus anything that successfully solves Equation 9.18 should generate quality temporal vector representations. Note that the update requires calculating inverse of a matrix (Refer Equation 9.13). This step is computationally expensive. Thus, to overcome this, we adopt an iterative approach (See below) similar to [225] and obtain our solution.

$$\mathbf{S}(t) = (1-\alpha)\sum_{b=1}^{B}(\alpha\mathbf{B}(t))^{b-1} \tag{9.19}$$

where B refers to the number of iterations. Once the iterative algorithm converges, we can obtain our time-aware embeddings as $\mathbf{V}'(t) = \mathbf{X}'t+\mathbf{Y}'$. As our vector representations are parameterized with time, it allows us to predict the future co-occurrence matrix $\mathbf{Y}(t+1) \approx \mathbf{U}\mathbf{V}'(t+1)^T$. The entry values in $\mathbf{Y}(t+1)$ quantify the likelihood of future association (hypothesis) between biomedical concepts. Now, given an input concept of interest (A), the candidate concepts (C) are ranked based on their predicted future co-occurrence value and then presented to the user for further analysis and investigation. Having described the nuances of our methodology, in the next section we describe our experimental protocol and perform extensive analysis to validate the effectiveness of proposed approach.

9.4 Experiments

In this section, we demonstrate the efficacy of our proposed framework. Towards this end, we perform both qualitative and quantitative evaluations. The qualitative evaluation determines the extent to which our approach is capable of rediscovering the already known knowledge (and potentially new knowledge), whereas the quantitative evaluation is intended to analyze the overall quality of predictions/discoveries made by the system.

Dataset Description: MEDLINE[3] , the largest available scientific repository, is used as the primary source of information for performing experiments. At present, it provides access to more than 24 million time-stamped articles primarily from the domain of life-sciences and bio-medicine. Among others, each article in MEDLINE contains the following attributes: a) unique identifier known as PMID, b) title, c) abstract, d) publication date and e) Medical Subject Headings (MeSH) terms. Previous studies [9] have shown that using concepts from raw title/abstract may introduce noise to the system and prove computationally expensive. To circumvent this problem, a majority of studies [6, 208, 196] conduct their investigation studies by choosing MeSH terms as their unit of analysis. MeSH terms in MEDLINE refer to a set of special keywords that are assigned to each article by the subject-matter-experts. As the experts annotate these terms based on the full-content of the article, they can be assumed to represent the conceptual meaning of an article. Being manually curated, they are highly accurate and find their utility in a multitude of downstream biomedical applications. Considering its high input quality and broader applicability, in this study, we use MeSH terms as our unit of analysis[4] . Fortunately, these MeSH terms are also arranged in a hierarchical/taxonomic structure[5] . In our study, this taxonomic structure of MeSH terms serve as our Knowledge-base. As of year 2018, there are approximately 28,000 MeSH terms (V). For our experiments, we generate the temporal embeddings for these medical concepts. As recommended in some of the prior studies [16, 77], we set the dimensionality of our temporal embeddings to $n = 200$. The hyper-parameter for exponential decay function is set to $\theta = 0.3$. The regularization weights $\beta = \gamma_1 = \gamma_2 = \omega_1 = \omega_2 =$ is 0.01. The value of α in Equation 9.11 is empirically set to 0.5. Finally, the number of iteration for model and the value of B in Equation 9.19 are both set to 200.

[3] https://www.nlm.nih.gov/bsd/medline.html
[4] https://github.com/kishlayjha/hypotheses-generation-coEvolution
[5] https://www.nlm.nih.gov/mesh/intro_trees.html

9.4.1 Qualitative evaluation

To perform qualitative assessment, we borrow experimental settings from the hypotheses generation literature [6, 208]. A common way of performing evaluation is to replicate the five golden test-cases (enumerated below) reported by the pioneers in this area of study. For the sake of uniformity, we adopt the same setting and run the proposed model on these test-cases and probe for the results.

1. Raynaud's Disease (RD) and Fish Oils (FO) (**1985**)
2. Migraine Disorder (MIG) and Magnesium (MG) (**1988**)
3. Arginine (ARG) and Somatomedin C (IGF1) (**1994**)
4. Alzheimer Disease (AD) Indomethacin (INN) (**1989**)
5. Schizophrenia (SZ) and Calcium - Independent Phospholipase A2 (PA2) (**1997**)

To recapitulate our problem statement, the input to our hypothesis generation algorithm is a topic of interest (A) (e.g., Raynaud's disease), date (d) (e.g., 1985) and the goal is to find new biological relationships (C) (e.g., Fish Oils). The date (d) in the input acts as a cut-off threshold. Both the proposed model and baseline algorithms are run on the *pre-cut-off* segment (before date d) and the obtained results (predicted connections) are evaluated in the *post-cut-off* segment (after date d). To analyze the predicted results, we need a ground truth. However, there is no standard ground truth available and creating one remains an open problem [208]. Therefore, for the purpose of quantitative analysis, a supposedly ground truth is constructed. All those connections that co-occur with the input concept of interest in the post-cut-off segment but not in the pre-cut-off segment are assumed to be valid connections. These valid connections are ranked based on their TF-IDF co-occurrence score with the input concept of interest. The candidate set for target 'C' terms are all the concepts present in vocabulary besides - A and *Co-occur(A)*. *Co-occur(A)* refers to the set of terms that have co-occurred with A before the threshold date d. All the possible target terms are ranked based on their predicted co-occurrence score with the input concept of interest. Then, the top-k results are presented to the user. Semantic filters are needed because in the biomedical domain practitioners have a diverse range of interest. Some experts working in a specific area (ex: Genes or Drugs) might be interested only in those terms that have a possible genetic linkages or posses certain chemical properties. On the other hand, a novice biomedical scientist might have a general interest and is possibly looking for a surprising (or radical) connection. To emphasize our focus on finding potential

therapeutic preventions (and in the interest of space), we report results only for the semantic category "Drugs". Now, in the rest of this section, we discuss the ability of proposed model to rediscover the already known knowledge.

Raynaud's Disease (RD) and Fish Oils (FO): To replicate this knowledge, we seeded our HG system with input concept (A) as "Raynaud disease" and a date (d) as "1985". The objective is to find possible treatments (e.g, "Fish Oils") or other terms of biological significance in the top-k results. The top-k results for this and all other test cases are reported in Table, along with the evidences in the form of PMIDS. As it can be observed, the target term "Fish Oils" in ranked 3. If we filter the terms by Semantic category "Drug", the term "Fish Oils" obtain rank 1.

Migraine Disorder (MIG) and Magnesium (MG): In 1988, the authors in [78] studied the possible linkage between "Migraine Disorder" and "Magnesium". In their conclusion, the authors reported eleven previously unknown connections. In our results, we found the target term at rank 5 (overall) and rank 2 (semantic filter - Drug) respectively.

Arginine (ARG) and Somatomedin C (IGF1): In this test-case, the authors [8] explored the relationship between a growth-regulating peptide (i.e., Somatomedin C) and an amino acid (i.e., Arginine). In our results, we found the target concept *Somatomedin C* at rank 2 (overall) and 3 (semantic filter - Drug) respectively.

Alzheimer Disease (AD) Indomethacin (INN): The objective of this case-study was to find a possible connection between Indomethacin (an anti-inflammatory agent) and Alzheimer Disease (a progressive disorder that cause memory loss and other mental issues) [8]. The target term "Indomethacin" is ranked 5 (Overall) and 2 (Semantic filter - Drug) respectively.

Schizophrenia (SZ) and Calcium - Independent Phospholipase A2 (PA2): Schizophrenia is a chronic disorder that affects person's ability to think, feel and reason clearly [8]. In our results, the target term Phospholipase A2 (PA2) was ranked 3 (Overall) and 2 (Semantic - Filter) respectively.

Discovery example for the case of Autism: In our experiments, we tried to analyze the results of proposed approach on new test cases. To do so, we choose a disease of biomedical significance: *Autism*. Autism is a serious development disorder found in children that impairs the ability to communicate and interact. We seeded our algorithm with input as "Autism", date (d) as "2014" and analyzed the top-k results. The top term found was "calcineurin" (a protein phosphate). Upon manually inspecting the medical

literature, we found that there might exist an indirect link between the calcineurin and autism via terms such as "Bcl-2", "calmodulin" and "synaptic plasticity". Although clinical trails are needed to corroborate any hypothesis, several recent studies [226] suggest that these terms are of potential clinical interest.

From the results of above qualitative analysis, one can infer that the proposed HG system is able to successfully replicate the known knowledge and potentially discover new practical knowledge. While this form of evaluation provides insight into the quality of top-ranked results, a quantitative form of evaluation is necessary to gain an understanding of overall results.

9.4.2 Quantitative evaluation

The objective of this section is to examine the overall quality of prediction/discoveries generated. To achieve this, we split the corpus into pre-segment/post-segment (Refer Section 9.4.1), and obtain the ranked set for both generated connections and ground truth. Then, *Spearman coefficient* is used to measure the performance. As a post-processing step, all the trivial connections (check-tags [3] such as "humans", "male", "female" and so on) are removed from both the ground truth and predicted set. Next in this section, we report the quantitative results and discuss our findings on all the five test-cases enumerated in Section 9.4.1. In this regard, one might question: How is the performance of HG systems in test-cases other than the traditional five test-cases? To answer this, we choose 200 diseases of biomedical significance and conducted experiments using the same timeslicing scheme. Specifically, for each of these 200 diseases, we set the cut-off date to January 1, 2014, which resulted in a pre-cut-off set composed of 19,895,212 million documents published before January 1, 2014 and a post-cut-off set composed of 4,587,929 documents published after January 1, 2014. The results obtained are reported and analyzed later in this section.

Evaluation baselines for quantitative evaluation: To compare the performance of proposed model with existing hypothesis generation systems, the following six baseline algorithms are implemented.

1. *Jaccard*: Jaccard is a popular link prediction technique. The formula to calculate the strength of association between two concepts is given below:

   $Association(A, C) = \mid Count_A \cap Count_C \mid / \mid Count_A \cup Count_C \mid$, where $Count_i$ refers the set of terms that co-occur with i.

2. *Preferential Attachment*: Preferential Attachment is another classical link prediction technique. The formula to calculate preferential attachment is given below:

   $Association(A, C) =\mid Count_A \mid + \mid Count_C \mid$, where $Count_i$ refers the set of terms that co-occur with i.

3. *Arrowsmith*: Arrowsmith is a popular hypothesis generation system proposed in [8].

4. *BITOLA*: BITOLA is a recent hypothesis generation algorithm proposed in [227].

5. *Static Embeddings*: Static embeddings refers to the word embeddings generated from given corpus without incorporating any temporal component. The static embeddings are generated by training the standard CBOW [16] model on the entire MEDLINE corpus. All the hyper-parameters for CBOW are chosen as suggested in the study [16].

6. *Dynamic MeSH Embedding* [23]: DME refers to a recent HG algorithm that models the semantic evolution of medical concepts from the diachornic biomedical corpora alone. It does not incorporate the (co)-evolving features of medical concepts from contemporary knowledge bases.

Note that the first two algorithms (Jaccard and Preferential Attachment) are from the link prediction literature. As we formulated the current task into a weighted link prediction problem, it is of interest to compare the results with classical link prediction techniques.

Evaluation metrics for quantitative evaluation: Two evaluation metrics are used to quantify our results: 1) Spearman Coefficient@k and 2) Mean Average Precision (MAP@k).

Results: Table 9.1, 9.2, 9.3, 9.4, 9.5 reports the Spearman-Coefficient@k for each of the five golden datasets enumerated in Section 9.4.1. The value of K is gradually increased from top 200 to 1500 and results are reported. Table 9.6 reports the MAP@K by consolidating numbers across 200 diseases (excluding the five golden test-cases) of biomedical significance.

Discussion: From Tables 9.1, 9.2, 9.3, 9.4, 9.5 and 9.6 it can be observed that the proposed model consistently outperforms all the existing baselines in terms of both Spearman-Coefficient@K and MAP@K. This result indicates the ability of proposed framework to find semantically meaningful connections at top ranks. Analyzing the

Table 9.1: Spearman's Correlation for FO-RD

Algorithm	k=200	k=800	k=1000	k=1500
Jaccard	0.012	0.011	0.017	0.102
Preferntial attachment	0.004	0.006	0.009	0.101
Arrowsmith	0.018	0.013	0.012	0.106
BITOLA	0.019	0.021	0.018	0.119
Static (No evolution)	0.027	0.031	0.019	0.127
DME (No co-evolution)	0.068	0.081	0.101	0.189
Proposed	**0.189**	**0.205**	**0.301**	**0.407**

Table 9.2: Spearman's Correlation for MG-MIG

Algorithm	k=200	k=800	k=1000	k=1500
Jaccard	0.017	0.023	0.009	0.109
Preferntial attachment	0.019	0.026	0.011	0.112
Arrowsmith	0.021	0.041	0.017	0.115
BITOLA	0.023	0.042	0.019	0.127
Static (No evolution)	0.034	0.061	0.027	0.136
DME (No co-evolution)	0.078	0.092	0.109	0.193
Proposed	**0.179**	**0.275**	**0.389**	**0.469**

Table 9.3: Spearman's Correlation for AD-INN

Algorithm	k=200	k=800	k=1000	k=1500
Jaccard	0.012	0.014	0.018	0.100
Preferential attachment	0.011	0.013	0.017	0.112
Arrowsmith	0.014	0.023	0.038	0.118
BITOLA	0.027	0.032	0.047	0.124
Static (No evolution)	0.036	0.045	0.101	0.137
DME (No co-evolution)	0.058	0.079	0.112	0.187
Proposed	**0.197**	**0.292**	**0.362**	**0.447**

Table 9.4: Spearman's Correlation for IGF1-ARG

Algorithm	k=200	k=800	k=1000	k=1500
Jaccard	0.018	0.026	0.013	0.101
Preferntial attachment	0.022	0.012	0.017	0.103
Arrowsmith	0.022	0.031	0.017	0.104
BITOLA	0.026	0.032	0.018	0.119
Static (No evolution)	0.033	0.082	0.028	0.157
DME (No co-evolution)	0.092	0.097	0.125	0.194
Proposed	**0.280**	**0.385**	**0.425**	**0.487**

Table 9.5: Spearman's Correlation for SZ-PA2

Algorithm	k=200	k=800	k=1000	k=1500
Jaccard	0.024	0.014	0.095	0.112
Preferntial attachment	0.023	0.015	0.017	0.121
Arrowsmith	0.089	0.029	0.102	0.136
BITOLA	0.092	0.031	0.108	0.143
Static (No evolution)	0.017	0.095	0.129	0.195
DME (No co-evolution)	0.098	0.164	0.157	0.278
Proposed	**0.187**	**0.224**	**0.384**	**0.416**

Table 9.6: Mean Average Precision@k for 200 disease

Algorithm	k=200	k=800	k=1000	k=1500
Jaccard	0.012	0.013	0.017	0.102
Preferntial attachment	0.011	0.012	0.015	0.103
Arrowsmith	0.018	0.011	0.012	0.106
BITOLA	0.019	0.021	0.018	0.119
Static (No evolution)	0.027	0.031	0.019	0.127
DME (No co-evolution)	0.068	0.081	0.101	0.189
Proposed	**0.185**	**0.262**	**0.392**	**0.435**

overall results from different perspectives, we detect various trends. First, the contemporary HG systems - ARROWSMITH and BITOLA - perform better than classical link prediction techniques. This highlights the challenges that are unique to HG task and encourages us to develop solutions tailored to HG. Second, we notice that though the contemporary HG algorithms perform better than link prediction techniques, they fall behind the Static embedding approach. Upon manual inspection of results, we found that this is mainly due to two factors: a) over reliance on co-occurrence statistics, b) failing to capture the implicit semantics of medical concepts. To elaborate, the baseline HG algorithms (Number 3 and 4) are purely distributional in nature. This results in promoting those terms that are "contextually generic". Contextually generic terms are those terms that co-occur frequently with the input concept of interest but have meager semantic meaning associated to them. For instance, consider the example of "Migraine Disorder". Some of the related terms that frequently co-occur with Migraine are "headache", "pain". While these terms are statistically associated to "Migraine", they have poor semantic association. As baseline HG algorithms rely strongly on statistical co-occurrence, these contextually generic terms are ranked higher. This proves counter-intuitive as these same terms are ranked lower in the ground truth. Another point we wish to highlight is that, as embeddings based approaches are capable of capturing the implicit semantics, they successfully promote those terms that have functional relationship with input concept of interest. Recall that the word embeddings can capture special features such as linear analogical relationships *vec("ibuprofen") - vec("headache")* $\approx$ *vec("treats")*. This special feature provides leverage to embedding based techniques over other approaches. Third, we observe that a recent temporal embedding based approach [23] performs better than Static embedding [16]. This result highlights the importance of leveraging the semantic change of concepts for predictive tasks such as Hypothesis generation. Lastly, we would like to highlight that the proposed model outperforms the existing temporal embedding approach. This is because the existing temporal embedding based approach [23] fails to leverage the (co)-evolutionary features of medical concepts from contemporary KB's. In our experiments, we found that such subject-matter-expert maintained KB's have invaluable information and their incorporation is important to generate robust temporal embeddings. Furthermore, we noticed that the collaborative exploitation of semantics from natural language text and KB's proved particularly helpful for domain-specific (rare) words. As an illustration,

consider the medical concept "Adioisotopes". This concept rarely co-occurs with "Magnesium" but is known to have strong semantic association with it. The recent temporal word embedding approach [23] (without external knowledge) fails to identify this term (and such domain-specific words in general) in top-ranks, due to the lack of sufficient statistical information. While such domain-specific words lack local-context information, their semantics can be mined from human curated KB's. As the proposed framework effectively leverages the KB's, such domain-specific terms are successfully promoted to higher ranks in our predicted set, thereby resulting in improved performance. In summary, from our both qualitative and quantitative experiments, we conclude that jointly leveraging the local-context information from natural language text and topological features from knowledge-base aids to generate temporal embeddings that are both robust and posses better predictive power, thereby, generating effective hypothesis.

9.5 Conclusions

In this study, we proposed a general framework for hypothesis generation that models the temporal (co)-evolution of biomedial concepts from two complementary sources of information - corpus and domain knowledge. By synthesizing the mutual evolution of concepts from these intertwined resources, the proposed model generates temporal embeddings that are both robust and posses higher predictive effects. Technically, the model achieves this by adopting a temporal co-factorization framework wherein the sub-spaces between multiple related matrices are learned by sharing a constant factor. Both qualitative and quantitative experiments conducted on the largest biomedical corpora validates the efficacy of the proposed approach, and suggests that the proposed framework has potential for generating new practical knowledge.

Chapter 10

Conclusions and Future Directions

An advanced hypothesis generation system that generates "actionable" postulates is a particularly difficult task, given the intrinsic complexities present in the process of imitating the steps a cognitive mind undertakes while forging a plausible hypothesis. However, the massive amount of data being generated by health-care sector and its current trend towards rapid digitization has overwhelmed the domain experts. Consequently, it has become necessary to design a system that can process, analyze this quintessential (bio)-medical "big data" and generate promising hypotheses that could be potentially validated, thereby, benefiting the society at large.

In this direction, the proposed framework is our initial step towards developing a robust hypothesis generation system. Aside from the evolutionary characteristics it carries and the flexibility it provides in integrating heterogeneous textual sources, it is computationally tractable. This allows the end users to obtain the desired results in a reasonable amount time. Furthermore, another crucial advantage of this framework lies in its ability to present evolution trajectory visualization of medical concepts. For any two medical concepts, their semantic progression over time can be analyzed and understood. This form of visualization is believed to aid domain experts in making informed choices. Finally, the system also provides evidences (PubMED article identifier) for its outputs thus making its results interpretable.

The results indicate that the hypotheses generated by the proposed framework are encouraging. As an illustration, consider the bridge concept *nitric oxide* that was found

for the test-case of Alzheimer's disease and Indomethacin. Although not clinically corroborated yet, several papers identified nitric oxide as an important element for understanding Alzheimer's Disease [198]. Additionally, during 2000-2001, there were studies [79] showing evidence of strong influence of nitric oxide in both Alzheimer's disease and Indomethacin. Thus, the results suggest that the proposed methodology is capable of generating both semantically meaningful and temporally sensible hypotheses that are worthy of clinical trails and further investigation.

In our continuing research, we are investigating in several directions. First is to speculate more sophisticated approaches to generate medical concept embeddings that are able to capture the multifaceted aspects of semantic expressiveness. Another area of interest is to explore the application of our methodology to tasks such as drug-drug interaction, adverse-drug events and biomedical question answering.

References

[1] Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. Probing biomedical embeddings from language models. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 82–89, 2019.

[2] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[3] Zhiyong Lu. Pubmed and beyond: a survey of web tools for searching biomedical literature. *Database*, 2011:baq036, 2011.

[4] Doug Howe, Maria Costanzo, Petra Fey, Takashi Gojobori, Linda Hannick, Winston Hide, David P Hill, Renate Kania, Mary Schaeffer, Susan St Pierre, et al. Big data: The future of biocuration. *Nature*, 455(7209):47–50, 2008.

[5] Don R Swanson. Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1):7–18, 1986.

[6] Padmini Srinivasan. Text mining: Generating hypotheses from medline. *J. Assoc. Inf. Sci. Technol.*, 55(5):396–413, 2004.

[7] Kishlay Jha, Guangxu xun, Yaqing Wang, Vishrawas Gopalakrishnan, and Aidong Zhang. Concepts-bridges: Uncovering conceptual bridges based on biomedical concept evolution. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 1599–1607, New York, NY, USA, 2018. ACM.

[8] Don R Swanson and Neil R Smalheiser. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial intelligence*, 91(2):183–203, 1997.

[9] Meliha Yetisgen-Yildiz and Wanda Pratt. Using statistical and knowledge-based approaches for literature-based discovery. *Journal of biomedical informatics*, 39(6):600–611, 2006.

[10] D. Cameron, R. Kavuluru, T. C. Rindflesch, A. P. Sheth, K. Thirunarayan, and O. Bodenreider. Context-driven automatic subgraph creation for literature-based discovery. *J Biomed Inform*, 54:141–57, Apr 2015.

[11] B. Wilkowski, M. Fiszman, C. M. Miller, D. Hristovski, S. Arabandi, G. Rosemblat, and T. C. Rindflesch. Graph-based methods for discovery browsing with semantic predications. *AMIA Annu Symp Proc*, 2011:1514–23, 2011.

[12] D. Weissenborn, M. Schroeder, and G. Tsatsaronis. Discovering relations between indirectly connected biomedical concepts. *J Biomed Semantics*, 6:28, 2015.

[13] Anika Groß, Cédric Pruski, and Erhard Rahm. Evolution of biomedical ontologies and mappings: Overview of recent approaches. *Computational and structural biotechnology journal*, 14:333–340, 2016.

[14] T. C. Rindflesch and M. Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J Biomed Inform*, 36(6):462–77, Dec 2003.

[15] Caroline B Ahlers, Marcelo Fiszman, Dina Demner-Fushman, François-Michel Lang, and Thomas C Rindflesch. Extracting semantic predications from medline citations for pharmacogenomics. In *Biocomputing 2007*, pages 209–220. World Scientific, 2007.

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

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

[18] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[19] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *journal of the Association for Information Science and Technology*, 58(7):1019–1031, 2007.

[20] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, 2011.

[21] Xiaoyi Li, Nan Du, Hui Li, Kang Li, Jing Gao, and Aidong Zhang. A deep learning approach to link prediction in dynamic networks. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 289–297. SIAM, 2014.

[22] Vishrawas Gopalakrishnan, Kishlay Jha, Guangxu Xun, Hung Q Ngo, and Aidong Zhang. Towards self-learning based hypotheses generation in biomedical text domain. *Bioinformatics*, 34(12):2103–2115, 2017.

[23] Guangxu Xun, Kishlay Jha, Vishrawas Gopalakrishnan, Yaliang Li, and Aidong Zhang. Generating medical hypotheses based on evolutionary medical concepts. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 535–544. IEEE, 2017.

[24] 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.

[25] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

[26] Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain. Neural probabilistic language models. In *Innovations in Machine Learning*, pages 137–186. Springer, 2006.

[27] Thorsten Brants, Ashok C Popat, Peng Xu, Franz J Och, and Jeffrey Dean. Large language models in machine translation. In *In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Citeseer, 2007.

[28] Ying Zhang, Almut Silja Hildebrand, and Stephan Vogel. Distributed language modeling for n-best list re-ranking. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 216–223. Association for Computational Linguistics, 2006.

[29] Andriy Mnih and Koray Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in neural information processing systems*, pages 2265–2273, 2013.

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

[31] Chang Xu, Yalong Bai, Jiang Bian, Bin Gao, Gang Wang, Xiaoguang Liu, and Tie-Yan Liu. Rc-net: A general framework for incorporating knowledge into word representations. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1219–1228. ACM, 2014.

[32] Hsin-Yang Wang and Wei-Yun Ma. Integrating semantic knowledge into lexical embeddings based on information content measurement. *EACL 2017*, page 509, 2017.

[33] Quan Liu, Hui Jiang, Si Wei, Zhen-Hua Ling, and Yu Hu. Learning semantic word embeddings based on ordinal knowledge constraints. *Proceedings of ACL, Beijing, China*, 2015.

[34] Zhiguo Yu, Trevor Cohen, Elmer V Bernstam, and Byron C Wallace. Retrofitting word vectors of mesh terms to improve semantic similarity measures. *EMNLP 2016*, page 43, 2016.

[35] Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*, 2014.

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

[37] Mo Yu and Mark Dredze. Improving lexical embeddings with semantic knowledge. In *ACL (2)*, pages 545–550, 2014.

[38] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185, 2014.

[39] Hoa A Nguyen and Hoa Al-Mubaid. New ontology-based semantic similarity measure for the biomedical domain. In *Granular Computing, 2006 IEEE International Conference on*, pages 623–628. IEEE, 2006.

[40] Hisham Al-Mubaid and Hoa A Nguyen. Measuring semantic similarity between biomedical concepts within multiple ontologies. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 39(4):389–398, 2009.

[41] Mohamed Ben Aouicha and Mohamed Ali Hadj Taieb. Computing semantic similarity between biomedical concepts using new information content approach. *Journal of biomedical informatics*, 59:258–275, 2016.

[42] Ted Pedersen, Serguei VS Pakhomov, Siddharth Patwardhan, and Christopher G Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics*, 40(3):288–299, 2007.

[43] Angelos Hliaoutakis. Semantic similarity measures in mesh ontology and their application to information retrieval on medline. *Master's thesis*, 2005.

[44] 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 *AMIA annual symposium proceedings*, volume 2010, page 572. American Medical Informatics Association, 2010.

[45] Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on systems, man, and cybernetics*, 19(1):17–30, 1989.

[46] Claudia Leacock and Martin Chodorow. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283, 1998.

[47] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.

[48] Ikkyu Choi and Minkoo Kim. Topic distillation using hierarchy concept tree. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 371–372. ACM, 2003.

[49] SPFGH Moen and Tapio Salakoski2 Sophia Ananiadou. Distributional semantics resources for biomedical text processing, 2013.

[50] Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. How to train good word embeddings for biomedical nlp. *Proceedings of BioNLP16*, page 166, 2016.

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

[52] Serguei VS Pakhomov, Greg Finley, Reed McEwan, Yan Wang, and Genevieve B Melton. Corpus domain effects on distributional semantic modeling of medical terms. *Bioinformatics*, 32(23):3635–3644, 2016.

[53] Bridget T McInnes and Ted Pedersen. Improving correlation with human judgments by integrating semantic similarity with second–order vectors. *BioNLP 2017*, page 107, 2017.

[54] Montserrat Batet, David Sánchez, and Aida Valls. An ontology-based measure to compute semantic similarity in biomedicine. *Journal of biomedical informatics*, 44(1):118–125, 2011.

[55] Yuan Ling, Yuan An, Mengwen Liu, Sadid A Hasan, Yetian Fan, and Xiaohua Hu. Integrating extra knowledge into word embedding models for biomedical nlp tasks. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 968–975. IEEE, 2017.

[56] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

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

[58] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, Christopher Potts, et al. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642, 2013.

[59] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Hlt-naacl*, volume 13, pages 746–751, 2013.

[60] Ye Yuan, Guangxu Xun, Kebin Jia, and Aidong Zhang. A multi-view deep learning method for epileptic seizure detection using short-time fourier transform. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB 2017, Boston, MA, USA, August 20-23, 2017*, pages 213–222, 2017.

[61] Guangxu Xun, Yaliang Li, Jing Gao, and Aidong Zhang. Collaboratively improving topic discovery and word embeddings by coordinating global and local contexts. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 535–543, 2017.

[62] Yongjun Zhu, Erjia Yan, and Fei Wang. Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. *BMC Medical Informatics and Decision Making*, 17(1):95, 2017.

[63] Zhenchao Jiang, Lishuang Li, Degen Huang, and Liuke Jin. Training word embeddings for deep learning in biomedical text mining tasks. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, pages 625–628. IEEE, 2015.

[64] Asli Celikyilmaz, Dilek Hakkani-Tur, Panupong Pasupat, and Ruhi Sarikaya. Enriching word embeddings using knowledge graph for semantic tagging in conversational dialog systems. *genre*, 2010.

[65] Kishlay Jha, Guangxu Xun, Vishrawas Gopalakrishnan, and Aidong Zhang. Augmenting word embeddings through external knowledge-base for biomedical application. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 1965–1974. IEEE, 2017.

[66] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016.

[67] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. ACM, 2015.

[68] Yating Zhang, Adam Jatowt, Sourav S Bhowmick, and Katsumi Tanaka. The past is not a foreign country: Detecting semantically similar terms across time. *IEEE Transactions on Knowledge and Data Engineering*, 28(10):2793–2807, 2016.

[69] Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the eleventh acm international conference on web search and data mining*, pages 673–681, 2018.

[70] Robert Bamler and Stephan Mandt. Dynamic word embeddings. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 380–389, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[71] George Tsatsaronis, Iraklis Varlamis, Nattiya Kanhabua, and Kjetil Nørvåg. Temporal classifiers for predicting the expansion of medical subject headings. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 98–113. Springer, 2013.

[72] Sofia J Athenikos and Hyoil Han. Biomedical question answering: A survey. *Computer methods and programs in biomedicine*, 99(1):1–24, 2010.

[73] Wanda Pratt and Meliha Yetisgen-Yildiz. Litlinker: capturing connections across the biomedical literature. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 105–12, 2003.

[74] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[75] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[76] Shengwen Peng, Ronghui You, Hongning Wang, Chengxiang Zhai, Hiroshi Mamitsuka, and Shanfeng Zhu. Deepmesh: deep semantic representation for improving large-scale mesh indexing. *Bioinformatics*, 32(12):70–79, 2016.

[77] 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.

[78] Don R Swanson. Migraine and magnesium: eleven neglected connections. *Perspectives in biology and medicine*, 31(4):526–557, 1988.

[79] Ronen Feldman and Haym Hirsh. Exploiting background information in knowledge discovery from text. *Journal of Intelligent Information Systems*, 9(1):83–97, 1997.

[80] Brian M Ross, Craig Hudson, Jeffrey Erlich, Jerry J Warsh, and Stephen J Kish. Increased phospholipid breakdown in schizophrenia: evidence for the involvement of a calcium-independent phospholipase a2. *Archives of general psychiatry*, 54(5):487–494, 1997.

[81] Chia-Feng Kuo, Shun Cheng, and John R Burgess. Deficiency of vitamin e and selenium enhances calcium-independent phospholipase a2 activity in rat lung and liver. *The journal of nutrition*, 125(6):1419, 1995.

[82] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[83] V. Gopalakrishnan, K. Jha, A. Zhang, and W. Jin. Generating hypothesis: Using global and local features in graph to discover new knowledge from medical literature. In *BICOB*, pages 23–30, 2016.

[84] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[85] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[86] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[87] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[88] Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433. ACM, 2006.

[89] Chong Wang, David Blei, and David Heckerman. Continuous time dynamic topic models. *arXiv preprint arXiv:1206.3298*, 2012.

[90] Ye Yuan, Guangxu Xun, Qiuling Suo, Kebin Jia, and Aidong Zhang. Wave2vec: Learning deep representations for biosignals. In *Data Mining (ICDM), 2017 IEEE International Conference on*, pages 1159–1164. IEEE, 2017.

[91] Xin Liu, Duygu Tosun, Michael W Weiner, Norbert Schuff, Alzheimer's Disease Neuroimaging Initiative, et al. Locally linear embedding (lle) for mri based alzheimer's disease classification. *Neuroimage*, 83:148–157, 2013.

[92] Duc Luu Ngo, Naoki Yamamoto, Vu Anh Tran, Ngoc Giang Nguyen, Dau Phan, Favorisen Rosyking Lumbanraja, Mamoru Kubo, and Kenji Satou. Application of word embedding to drug repositioning. *Journal of Biomedical Science and Engineering*, 9(01):7, 2016.

[93] Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah Smith. Sparse overcomplete word vector representations. *arXiv preprint arXiv:1506.02004*, 2015.

[94] Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. Ultradense word embeddings by orthogonal transformation. *arXiv preprint arXiv:1602.07572*, 2016.

[95] Sungjoon Park, JinYeong Bak, and Alice Oh. Rotated word vector representations and their interpretability. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 401–411, 2017.

[96] Brian Murphy, Partha Talukdar, and Tom Mitchell. Learning effective and interpretable semantic models using non-negative sparse embedding. *Proceedings of COLING 2012*, pages 1933–1950, 2012.

[97] Hongyin Luo, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Online learning of interpretable word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1687–1692, 2015.

[98] Fei Sun, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. Sparse word embeddings using l1 regularized online learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2915–2921. AAAI Press, 2016.

[99] Gregory Murphy. *The big book of concepts*. MIT press, 2004.

[100] Alexa T McCray, Anita Burgun, and Olivier Bodenreider. Aggregating umls semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, 84(0 1):216, 2001.

[101] Pengwei Wang, Tianyong Hao, Jun Yan, and Lianwen Jin. Large-scale extraction of drug–disease pairs from the medical literature. *Journal of the Association for Information Science and Technology*, 68(11):2649–2661, 2017.

[102] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems*, pages 585–591, 2002.

[103] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM, 2016.

[104] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[105] Ming Gao, Xiangnan He, Leihui Chen, Tingting Liu, Jinglin Zhang, and Aoying Zhou. Learning vertex representations for bipartite networks. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[106] Elissa J Chesler and Michael A Langston. Combinatorial genetic regulatory network analysis tools for high throughput transcriptomic data. In *Systems biology and regulatory genomics*, pages 150–165. Springer, 2005.

[107] Erich J Baker, Jeremy J Jay, Vivek M Philip, Yun Zhang, Zuopan Li, Roumyana Kirova, Michael A Langston, and Elissa J Chesler. Ontological discovery environment: A system for integrating gene–phenotype associations. *Genomics*, 94(6):377–387, 2009.

[108] Zhiyuan Chen and Bing Liu. Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 10(3):1–145, 2016.

[109] Peng Cui, Xiao Wang, Jian Pei, and Wenwu Zhu. A survey on network embedding. *IEEE Transactions on Knowledge and Data Engineering*, 31(5):833–852, 2018.

[110] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.

[111] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, pages 1067–1077. International World Wide Web Conferences Steering Committee, 2015.

[112] Daixin Wang, Peng Cui, and Wenwu Zhu. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1225–1234. ACM, 2016.

[113] Palash Goyal, Sujit Rokka Chhetri, and Arquimedes Canedo. dyngraph2vec: Capturing network dynamics using dynamic graph representation learning. *Knowledge-Based Systems*, 187:104816, 2020.

[114] Nansu Zong, Hyeoneui Kim, Victoria Ngo, and Olivier Harismendy. Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations. *Bioinformatics*, 33(15):2337–2344, 2017.

[115] Kishlay Jha, Yaqing Wang, Guangxu Xun, and Aidong Zhang. Interpretable word embeddings for medical domain. In *2018 IEEE international conference on data mining (ICDM)*, pages 1061–1066. IEEE, 2018.

[116] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. International Conference on Learning Representations, 2018.

[117] Seyed-Vahid Sanei-Mehri, Ahmet Erdem Sariyuce, and Srikanta Tirthapura. Butterfly counting in bipartite networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2150–2159. ACM, 2018.

[118] Yun Zhang, Charles A Phillips, Gary L Rogers, Erich J Baker, Elissa J Chesler, and Michael A Langston. On finding bicliques in bipartite graphs: a novel algorithm and its application to the integration of diverse biological data types. *BMC bioinformatics*, 15(1):110, 2014.

[119] Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. Thinking globally, acting locally: Distantly supervised global-to-local knowledge selection for background based conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8697–8704, 2020.

[120] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl_1):D535–D539, 2006.

[121] Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, 41(W1):W518–W522, 2013.

[122] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd*

ACM SIGKDD international conference on knowledge discovery and data mining, pages 135–144, 2017.

[123] Jainisha Sankhavara and Prasenjit Majumder. Biomedical information retrieval. In *FIRE (Working Notes)*, pages 154–157, 2017.

[124] Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):52, 2019.

[125] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3606–3611, 2019.

[126] Kyeonghye Guk, Gaon Han, Jaewoo Lim, Keunwon Jeong, Taejoon Kang, Eun-Kyung Lim, and Juyeon Jung. Evolution of wearable devices with real-time disease monitoring for personalized healthcare. *Nanomaterials*, 9(6):813, 2019.

[127] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[128] Bingbing Li, Zhenglun Kong, Tianyun Zhang, Ji Li, Zhengang Li, Hang Liu, and Caiwen Ding. Efficient transformer-based large scale language representations using hardware-friendly block structured pruning. *arXiv preprint arXiv:2009.08065*, 2020.

[129] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[130] Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations*, 2019.

[131] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *International Conference on Learning Representations*, 2018.

[132] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, 2019.

[133] Faiza Khan Khattak, Serena Jeblee, Chloé Pou-Prom, Mohamed Abdalla, Christopher Meaney, and Frank Rudzicz. A survey of word embeddings for clinical text. *Journal of Biomedical Informatics: X*, 4:100057, 2019.

[134] Kishlay Jha, Guangxu Xun, and Aidong Zhang. Continual representation learning for evolving biomedical bipartite networks. *Bioinformatics*, 37(15):2190–2197, 2021.

[135] Kishlay Jha, Guangxu Xun, Yaqing Wang, Vishrawas Gopalakrishnan, and Aidong Zhang. Concepts-bridges: Uncovering conceptual bridges based on biomedical concept evolution. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1599–1607, 2018.

[136] Kishlay Jha, Guangxu Xun, Yaqing Wang, and Aidong Zhang. Hypothesis generation from text based on co-evolution of biomedical concepts. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 843–851, 2019.

[137] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65, 2019.

[138] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*, 2020.

[139] Paul Ruvolo and Eric Eaton. Ella: An efficient lifelong learning algorithm. In *International Conference on Machine Learning*, pages 507–515, 2013.

[140] Guanyu Zhou, Kihyuk Sohn, and Honglak Lee. Online incremental feature learning with denoising autoencoders. In *Artificial intelligence and statistics*, pages 1453–1461, 2012.

[141] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

[142] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[143] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

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

[145] Margaret H Coletti and Howard L Bleich. Medical subject headings used to search the biomedical literature. *Journal of the American Medical Informatics Association*, 8(4):317–323, 2001.

[146] Laura Burdick, Jonathan K Kummerfeld, and Rada Mihalcea. Factors influencing the surprising instability of word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2092–2102, 2018.

[147] Bent Fuglede and Flemming Topsoe. Jensen-shannon divergence and hilbert space embedding. In *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE, 2004.

[148] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[149] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440, 2009.

[150] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10, 2014.

[151] 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(S2):S2, 2008.

[152] 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.

[153] María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920, 2013.

[154] Martin Krallinger, Obdulia Rabal, Saber A Akhondi, Martın Pérez Pérez, Jesús Santamaría, GP Rodríguez, 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.

[155] Yijia Zhang, Wei Zheng, Hongfei Lin, Jian Wang, Zhihao Yang, and Michel Dumontier. Drug–drug interaction extraction via hierarchical rnns on sequence and shortest dependency paths. *Bioinformatics*, 34(5):828–835, 2018.

[156] Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. Biosses: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58, 2017.

[157] Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. Medsts: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 54(1):57–72, 2020.

[158] Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, and Georgios Paliouras. Results of the seventh edition of the bioasq challenge. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 553–568. Springer, 2019.

[159] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, 2019.

[160] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Haotang Deng, and Qi Ju. Fastbert: a self-distilling bert with adaptive inference time. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 6035–6044, 2020.

[161] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. *arXiv preprint arXiv:1902.00751*, 2019.

[162] Huimin Luo, Min Li, Mengyun Yang, Fang-Xiang Wu, Yaohang Li, and Jianxin Wang. Biomedical data and computational models for drug repositioning: a comprehensive review. *Briefings in bioinformatics*, 2020.

[163] MM Malik, S Abdallah, and M Ala'raj. Data mining and predictive analytics applications for the delivery of healthcare services: a systematic literature review. *Annals of Operations Research*, 270(1):287–312, 2018.

[164] Kishlay Jha. Knowledge-base enriched word embeddings for biomedical domain. *arXiv preprint arXiv:2103.00479*, 2021.

[165] Anne Lauscher, Ivan Vulic, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. Informing unsupervised pretraining with external linguistic knowledge. 2019.

[166] Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. Sensebert: Driving some sense into bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667, 2020.

[167] Matthew E Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54, 2019.

[168] Christiane Fellbaum. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer, 2010.

[169] Kimberly J O'malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. Measuring diagnoses: Icd code accuracy. *Health services research*, 40(5p2):1620–1639, 2005.

[170] Ronald Cornet and Nicolette de Keizer. Forty years of snomed: a literature review. *BMC medical informatics and decision making*, 8(1):1–6, 2008.

[171] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6470–6479, 2017.

[172] Billy Chiu and Simon Baker. Word embeddings for biomedical natural language processing: A survey. *Language and Linguistics Compass*, 14(12):e12402, 2020.

[173] Bin He, Di Zhou, Jinghui Xiao, Xin Jiang, Qun Liu, Nicholas Jing Yuan, and Tong Xu. Integrating graph contextualized knowledge into pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2281–2290, 2020.

[174] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

[175] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[176] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 374–382, 2019.

[177] Guangxu Xun, Kishlay Jha, Ye Yuan, Yaqing Wang, and Aidong Zhang. Meshprobenet: a self-attentive probe net for mesh indexing. *Bioinformatics*, 35(19):3794–3802, 2019.

[178] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, 2014.

[179] Nigel Collier and Jin-Dong Kim. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78, 2004.

[180] Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC bioinformatics*, 16(1):1–17, 2015.

[181] Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, and Georgios Paliouras. Results of the seventh edition of the bioasq challenge. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 553—-568, 2019.

[182] Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, Georgios Paliouras, and Ioannis Kakadiaris. Results of the fifth edition of the bioasq challenge. In *BioNLP 2017*, pages 48–57, 2017.

[183] Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, 2006.

[184] Dongfang Xu, Zeyu Zhang, and Steven Bethard. A generate-and-rank framework with semantic type regularization for biomedical concept normalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8452–8464, 2020.

[185] Thomas C Rindflesch, Halil Kilicoglu, Marcelo Fiszman, Graciela Rosemblat, and Dongwook Shin. Semantic MEDLINE: An advanced information management application for biomedicine. *Inform. Serv. Use*, 31(1-2):15–21, 2011.

[186] Nut Limsopatham and Nigel Collier. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1014–1023, 2016.

[187] Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, et al. Data and systems for medication-related text classification and concept normalization from twitter: insights from the social media mining for health (smm4h)-2017 shared task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283, 2018.

[188] Yen-Fu Luo, Weiyi Sun, and Anna Rumshisky. Mcn: a comprehensive corpus for medical concept normalization. *Journal of biomedical informatics*, 92:103132, 2019.

[189] Zulfat Miftahutdinov and Elena Tutubalina. Deep neural models for medical concept normalization in user-generated texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 393–399, 2019.

[190] Jinghao Niu, Yehui Yang, Siheng Zhang, Zhengya Sun, and Wensheng Zhang. Multi-task character-level attentional networks for medical concept normalization. *Neural Processing Letters*, 49(3):1239–1256, 2019.

[191] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. Ernie: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, 2019.

[192] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908, 2020.

[193] Diana Sousa and Francisco M Couto. Biont: deep learning using multiple biomedical ontologies for relation extraction. *Advances in Information Retrieval*, 12036:367, 2020.

[194] Magdalena Biesialska, Katarzyna Biesialska, and Marta R Costa-jussà. Continual lifelong learning in natural language processing: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6523–6541, 2020.

[195] Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. Lamol: Language modeling for lifelong language learning. In *International Conference on Learning Representations*, 2019.

[196] Xiaohua Hu, Xiaodan Zhang, Illhoi Yoo, Xiaofeng Wang, and Jiali Feng. Mining hidden connections among biomedical concepts from disjoint biomedical literature sets through semantic-based association rule. *International Journal of Intelligent Systems*, 25(2):207–23, 2010.

[197] Shengtian Sang, Zhihao Yang, Zongyao Li, and Hongfei Lin. Supervised learning based hypothesis generation from biomedical literature. *BioMed research international*, 2015, 2015.

[198] P. Srinivasan and B. Libbus. Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics*, 20 Suppl 1:i290–96, Aug 2004.

[199] Scott Spangler, Angela D Wilkins, Benjamin J Bachman, Meena Nagarajan, Tajhal Dayaram, Peter Haas, Sam Regenbogen, Curtis R Pickering, Austin Comer, Jeffrey N Myers, et al. Automated hypothesis generation based on mining

scientific literature. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1877–1886. ACM, 2014.

[200] Yating Zhang, Adam Jatowt, Sourav Bhowmick, and Katsumi Tanaka. Omnia mutantur, nihil interit: Connecting past with present by finding corresponding terms across time. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 645–655, 2015.

[201] Xing Wei and W Bruce Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185. ACM, 2006.

[202] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004.

[203] Guangrong Li and Xiaodan Zhang. Mining biomedical knowledge using mutual information ABC. In *Granular Computing (GrC), 2011 IEEE International Conference on*, pages 848–50, 2011.

[204] Don R Swanson, Neil R Smalheiser, and Vetle I Torvik. Ranking indirect connections in literature-based discovery: The role of medical subject headings. *Journal of the Association for Information Science and Technology*, 57(11):1427–1439, 2006.

[205] Jake Lever, Sitanshu Gakkhar, Michael Gottlieb, Tahereh Rashnavadi, Santina Lin, Celia Siu, Maia Smith, Martin Jones, Martin Krzywinski, and Steven J Jones. A collaborative filtering based approach to biomedical knowledge discovery. *Bioinformatics*, 2017.

[206] Andreina Poggi Stefania Muti Giuseppe Bonapace Franco Argentati Claudio Cervini Ferdinando Silveri, Rossella De Angelis. Relative roles of endothelial cell damage and platelet activation in primary raynaud's phenomenon (rp) and rp secondary to systemic sclerosis. *Scandinavian journal of rheumatology*, 30(5):290–296, 2001.

[207] N Yi Mok, James Chadwick, Katherine AB Kellett, Eva Casas-Arce, Nigel M Hooper, A Peter Johnson, and Colin WG Fishwick. Discovery of biphenylacetamide-derived inhibitors of bace1 using de novo structure-based molecular design. *Journal of medicinal chemistry*, 56(5):1843–1852, 2013.

[208] Meliha Yetisgen-Yildiz and Wanda Pratt. A new evaluation methodology for literature-based discovery systems. *Journal of biomedical informatics*, 42(4):633–643, 2009.

[209] Guangxu Xun, Kishlay Jha, Ye Yuan, and Aidong Zhang. Topic discovery for biomedical corpus using mesh embeddings. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 1–4. IEEE, 2019.

[210] Haoyu Wang, Xuan Wang, Yaqing Wang, Guangxu Xun, Kishlay Jha, and Jing Gao. Interhg: an interpretable and accurate model for hypothesis generation. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1552–1557. IEEE, 2021.

[211] Kishlay Jha and Aidong Zhang. Continual knowledge infusion into pre-trained biomedical language models. *Bioinformatics*, 38(2):494–502, 2022.

[212] Peng Yan, Wei Jin, and Kishlay Jha. Discovering semantic relationships between concepts from medline. In *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*, pages 370–373. IEEE, 2016.

[213] Kishlay Jha, Guangxu Xun, Vishrawas Gopalakrishnan, and Aidong Zhang. Dwemed: Dynamic word embeddings for medical domain. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(2):1–21, 2019.

[214] Guangxu Xun, Kishlay Jha, and Aidong Zhang. Meshprobenet-p: improving large-scale mesh indexing with personalizable mesh probes. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(1):1–14, 2020.

[215] Yaqing Wang, Fenglong Ma, Haoyu Wang, Kishlay Jha, and Jing Gao. Multimodal emergent fake news detection via meta neural process networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3708–3716, 2021.

[216] Kishlay Jha and Wei Jin. Mining hidden knowledge from the counterterrorism dataset using graph-based approach. In *International Conference on Applications of Natural Language to Information Systems*, pages 310–317. Springer, 2016.

[217] Kishlay Jha and Wei Jin. Mining novel knowledge from biomedical literature using statistical measures and domain knowledge. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 317–326, 2016.

[218] Vishrawas Gopalakrishnan, Kishlay Jha, Aidong Zhang, and Wei Jin. Generating hypothesis: Using global and local features in graph to discover new knowledge from medical literature. In *Proceedings of the 8th International Conference on Bioinformatics and Computational Biology, BICOB*, volume 2016, pages 23–30, 2016.

[219] Guangxu Xun, Kishlay Jha, Jianhui Sun, and Aidong Zhang. Correlation networks for extreme multi-label text classification. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1074–1082, 2020.

[220] Vishrawas Gopalakrishnan, Kishlay Jha, Guangxu Xun, Hung Q Ngo, and Aidong Zhang. Towards self-learning based hypotheses generation in biomedical text domain. *Bioinformatics*, 34(12):2103–2115, 2018.

[221] Vishrawas Gopalakrishnan, Kishlay Jha, Wei Jin, and Aidong Zhang. A survey on literature based discovery approaches in biomedical domain. *Journal of biomedical informatics*, 93:103141, 2019.

[222] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857, 2018.

[223] Yakub Sebastian, Eu-Gene Siew, and Sylvester O Orimaye. Emerging approaches in literature-based discovery: techniques and performance review. *The Knowledge Engineering Review*, 32, 2017.

[224] Wenchao Yu, Charu C Aggarwal, and Wei Wang. Temporally factorized network modeling for evolutionary network analysis. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 455–464. ACM, 2017.

[225] Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. 11 label propagation and quadratic criterion. 2006.

[226] James Humble, Kazuhiro Hiratsuka, Haruo Kasai, and Taro Toyoizumi. Intrinsic spine dynamics are critical for recurrent network learning in models with and without autism spectrum disorder. *bioRxiv*, page 525980, 2019.

[227] D. Hristovski, C. Friedman, T. C. Rindflesch, and B. Peterlin. Exploiting semantic relations for literature-based discovery. *AMIA Annu Symp Proc*, pages 349–53, 2006.