Constructing Knowledge Graphs and Their Biomedical Applications

This manuscript (permalink) was automatically generated from greenelab/knowledge-graph-review@a6d4cf2 on February 27, 2020.

Authors

- **David Nicholson**
  0000-0003-0002-5761 · danich1
  Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania · Funded by GBMF4552 and T32 HG000046

- **Casey S. Greene**
  0000-0001-8713-9213 · cgreene · greenescientist
  Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania; Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation

Abstract

1. Give high level description of review as it pertains to knowledge graphs (creation and application)

Introduction

Knowledge graphs are a practical resource for many real world applications. They have been used in social medial mining to classify nodes [1] or to create a recommendation system [2]. Knowledge graphs have also been used to understand natural language via interpreting simple questions and using relational information to provide answers [3,4]. In a biomedical setting these graphs have been used to prioritize genes relevant to disease [5,6,7,8], perform drug repurposing [9] and identify drug-target interactions [10].

Despite their utility, precisely defining a knowledge graph is a difficult task because there are multiple conflicting definitions [11]. For this review, we define a knowledge graph as the following: a resource that integrates single or multiple sources of information into the form of a graph. This graph allows for the capacity to make semantic interpretation, continuously incorporate new information and uncover novel hidden knowledge through computational techniques and algorithms. Based on this definition resources like Hetionet [9] would be considered a knowledge graph. Hetionet integrates multiple sources of information into the form of a graph (example shown in Figure 1) and was used to derive novel information concerning unique drug treatments [9]. We do not consider databases like DISEASES [12] and DrugBank [13] to be knowledge graphs. These resources contain essential information, but do not represent their data in graph form.

Knowledge graphs are often constructed from manually curated databases [9,14,15,16]. These sources provide previously established information that can be incorporated into a graph. For example, a graph using DISEASES [12] as a resource would have genes and diseases as nodes, while edges would be added between nodes that have an association. This example shows a single type of relationship; however, there are graphs that use databases with multiple relationships. Other approaches have used natural language processing techniques to build knowledge graphs [17,18]. One example used a text mining system to extract sentences that indicated a protein interacting with another protein [19]. Once these sentences have been identified, they are incorporated as evidence for establishing edges in a knowledge graph.

In this review we describe various approaches for constructing and applying knowledge graphs in a biomedical setting. We discuss the pros and cons of constructing a knowledge graph via manually curated databases and via text mining systems. We also compare assorted approaches for applying knowledge graphs to solve biomedical problems. Lastly, we conclude on the practicality of knowledge graphs and point out future applications that have yet to be explored.

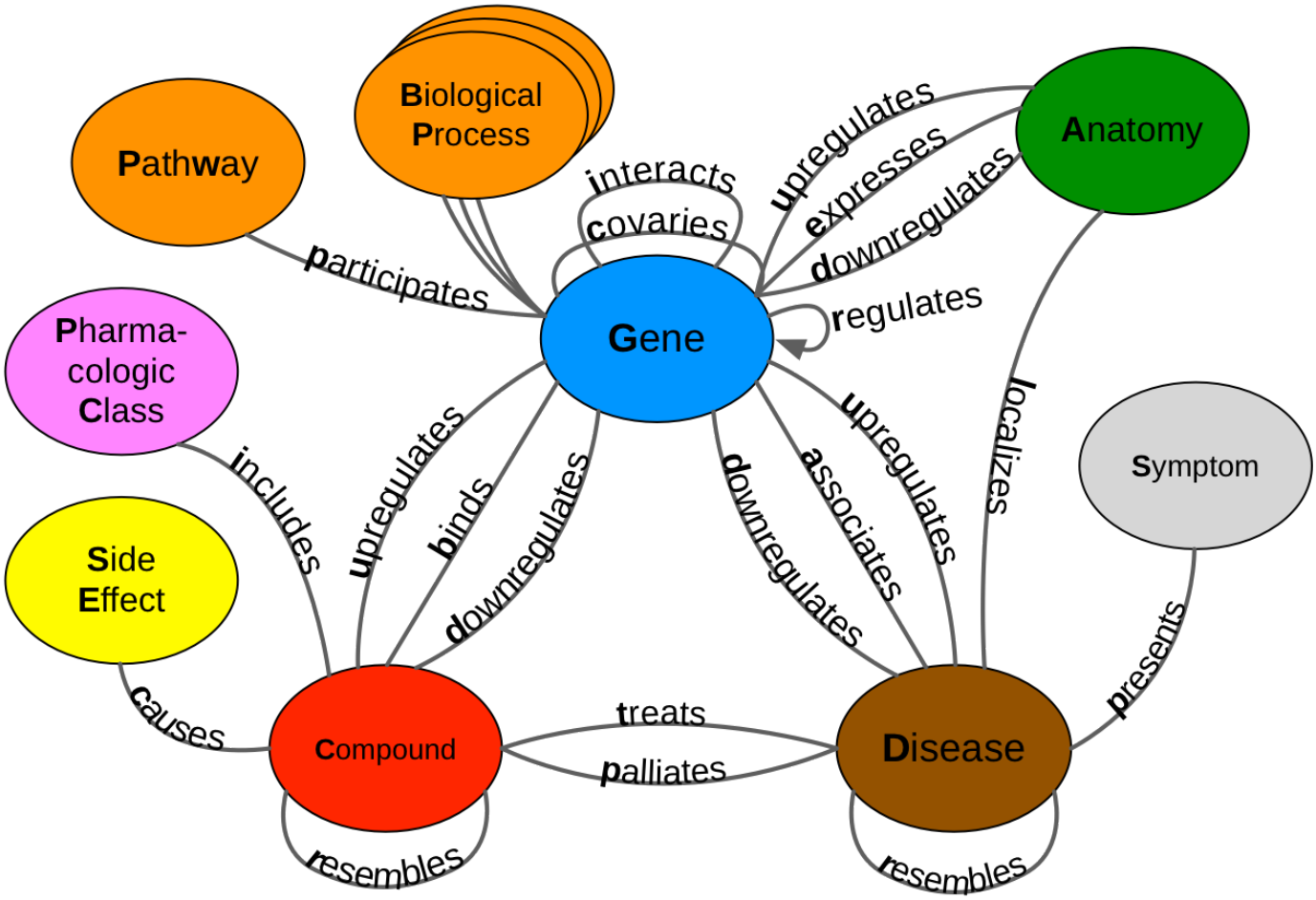

Figure 1: A metagraph (schema) of the heterogeneous network used in the Rephetio project [9]. This undirected network depicts pharmacological and biomedical information. The nodes (circles) represent entities and edges (lines) depict relational information between two entities.

Building Biomedical Knowledge Graphs

Knowledge graphs can be constructed in many ways using resources such as text or pre-existing databases. Usually, knowledge graphs are constructed using pre-existing databases. These databases are constructed by domain experts using approaches ranging from manual curation to automated techniques, such as text mining systems. Manual curation is a process that involves extensive use of domain experts to read papers and detect sentences that assert a relationship. Automated approaches involve the use of machine learning or natural language processing techniques to rapidly detect sentences of interest. We categorize these automated approaches into the following groups: rule-based extraction, unsupervised machine learning, and supervised machine learning. We discuss examples of each type of approach and synthesize the strengths and weaknesses of each.

Constructing Databases and Manual Curation

Database construction can date back all the way to 1956 where the first database contained a protein sequence of the insulin molecule [20]. This process involves gathering relevant text such as journal articles, abstracts, or web-based text. At this point curators can read gathered text and detect relationship asserting sentences (i.e. relationship extraction). An alternative to use a text mining system to filter out extraneous sentences, then incorporate curators to perfect the system's findings. This semi-automatic approach is way to augment curators throughout the curation process. We discuss the pros and cons of using manual curation for relationship extraction and mention databases that use this method to populate their fields.

Notable databases have been constructed via manual curation (Table {[**???**]}). For example, COSMIC [21] was constructed via a group of domain experts scanning the literature for key cancer related genes. This database has reached close to 35M entries in 2016 [21] and grew to a total of 45M entiries in 2019 [22]. Studies have shown that these databases contain relatively precise data, but in low quantities [23,24,25,26,27,28,29]. This happens because the high publication rate is too much for curators to keep up [30]. This findings highlight a critical need for future approaches to be fast enough to compete with an increasing publication rate.

Semi-automatic methods are a way to augment curators during the curation process [27,31,32,33,34,35,36]. First step in this context is to use an automatic system to initally extract sentences from text. This process filters out irrelevant sentences, which means less text for curators to sift through. After the pre-filtering step curators can approve or remove the identified sentences. This semi-automatic process was found to speed up the curation process compared to manual approach [31,37]. Curators in [37] saved an average of 2.8 hours of overall time while curators in [31] saved about the same amount of time (2 hours). Despite the speed up, this process is prone to produce bias results. As automated systems excel in identifying sentences for commonly occurring relationships, they miss out on lessor known relationships [31]. Plus, these systems have a hard time parsing ambiguous sentences that naturally occur in text. This complication results in curators have a difficult time correcting these systems [31]. Given these caveats, a future direction would be using or creating approaches that can mitigate the relationship bias. Furthermore, future approaches should look into using techniques that simplify sentences to solve the ambiguity issue [38,39].

Despite the negatives of manual curation, it is still an essential process for relationship extraction approaches. This process can be used to generate gold standard datasets that automated systems use for validation [40,41]. Furthermore, manual curation can be used during the training process of automated systems (i.e. active learning) [42]. It is important to remember that manual curation alone is precise, but results in low recall rates [29]. Future databases should consider initially relying on automated methods to obtain sentences at an acceptable recall level, then incorporate manual curation as a way to fix or remove irrelevant results.

Database [Reference]	Short Description	Number of Entries	Entity Types	Relationship Types	Method of Population
Entrez-Gene [43]	NCBI's Gene annotation database that contains information pertaining to genes, gene's organism source, phenotypes etc.	7,883,114	Genes, Species and Phenotypes	Gene-Phenotypes and Genes-Species mappings	Semi-automated curation
UniProt [44]	A protein protein interaction database that contains proteomic information.	560,823	Proteins, Protein sequences	Protein-Protein interactions	Manual and Automated Curation
PharmGKB [45]	A database that contains genetic, phenotypic, and clinical information related to pharmacogenomic studies.	43,112	Drugs, Genes, Phenotypes, Variants, Pathways	Gene-Phenotypes, Pathway-Drugs, Gene-Variants, Gene-Pathways	Manual Curation and Automated Methods

Database [Reference]	Short Description	Number of Entries	Entity Types	Relationship Types	Method of Population
COSMIC [21]	A database that contains high resolution human cancer genetic information.	35,946,704	Genes, Variants, Tumor Types	Gene-Variant Mappings	Manual Curation
BioGrid [46]	A database for major model organisms. It contains genetic and proteomic information.	572,084	Genes, Proteins	Protein-Protein interactions	Semi-automatic methods
Comparative Toxicogenomics Database [47]	A database that contains manually curated chemical-gene-disease interactions and relationships.	2,429,689	Chemicals (Drugs), Genes, Diseases	Drug-Genes, Drug-Disease, Disease-Gene mappings	Manual curation and Automated systems
Comprehensive Antibiotic Resistance Database [48]	Manually curated database that contains information about the molecular basis of antimicrobial resistance.	174,443	Drugs, Genes, Variants	Drug-Gene, Drug-Variant mappings	Manual curation
OMIM [49]	A database that contains phenotype and genotype information	25,153	Genes, Phenotypes	Gene-Phenotype mappings	Manual Curation

Table. A table of databases that used a form of manual curation to populate entries. Reported number of entities and relationships are relative to time of publication. {#tbl:manual-curated-databases}

Text Mining for Relationship Extraction

Rule-Based Relationship Extraction

Rule-based extraction consists of identifying sentences that contain important keywords or grammatical patterns that allude to relationships of interest. Keywords are established via expert knowledge or though the use of pre-existing ontologies. Grammatical patterns are constructed via experts curating parse trees, which are tree data structures that depict a sentence's grammatical structure. Parse trees come into two forms a constituency parse tree and a dependency parse trees. Both trees use part of speech tags, labels that dictate the grammatical role of a word such as noun, verb, adjective, etc, for construction. A constituency parse trees breaks a sentence down into a subphrases (Figure 3) while dependency path trees analyzes the grammatical structure of a sentence (Figure 2). Many text mining approaches [50,51,52] use such trees to generate features for machine learning algorithms. These approaches are discussed in later sections. For this section we focus on approaches that mainly use rule based extraction to detect sentences that assert a relationship.

Grammatical patterns can simplify sentences for easy extraction [39,53]. Jonnalagadda et al. used a set of grammar rules inspired by constituency trees to reshape complex sentences with simpler versions [39]. These simplified versions were manually curated to determine the presence of a relationship. By simplyfing sentences this approach achieved high recall, but had low precision [39]. Other approach used simplification techniques to make extraction easier [54,55,56,57]. Tudor et al., simplified sentences to detect protein phosphorylation events [56]. The sentence simplifier broke complex sentences that contain multiple protein events into smaller sentences that contain only one distinct event. By breaking these sentences down the authors were able to increase their recall. However, sentences that contained ambigious directionality or multiple phosphroylation events were too complex for the simplifier. As a consequence the simplifier produced errors in recall [56]. These errors highlight a crucial need for future algorithms to be generalizable enough to handle various forms of complex sentences.

Pattern matching is a fundamental approach used to detect relationship asserting sentences. In this context patterns can consist of phrases from constituency trees, a set of keywords or some combination of both to detect sentences [27,58,59,60,61,62]. Xu et al. designed a pattern matcher system to detect sentences in PubMed abstracts that indicate drug-disease treatments [61]. This system matched drug-disease pairs from clinicaltrails.gov to drug-disease pairs mentioned in abstracts. This matching process aided the authors in identifying sentences that were used to create simple patterns, such as “Drug in the treatment of Disease” [61], to match sentences in a wide variety of abstracts. The authors hand curated two datasets for evalution and achieved a high precision score of 0.904 and a low recall score of 0.131 [61]. This low recall score was based on constructed patterns being very specific to top occurring drug paris. This flaw resulted in rarely occurring pairs having a high likelihood of being missed. Following approaches using constituency trees, some approaches used dependency trees to construct patterns [50,63]. Depending upon the nature of the algorithm, dependency trees could be more appropiate than constituency trees and vise versa. The performance difference between the two approaches still remains as an open question for future exploration.

Rules based methods provide a basis for many relationship extraction systems. Approaches in this category range from simplifing sentences for easy extraction to identifing sentences based on matched key phrases or grammatical patterns. Both require a significant amount of manual effort and expert knowledge to perform well. A future direction is to develop ways to automatically construct these hand-crafted patterns, which would accelerate the process of creating new rule-based systems.

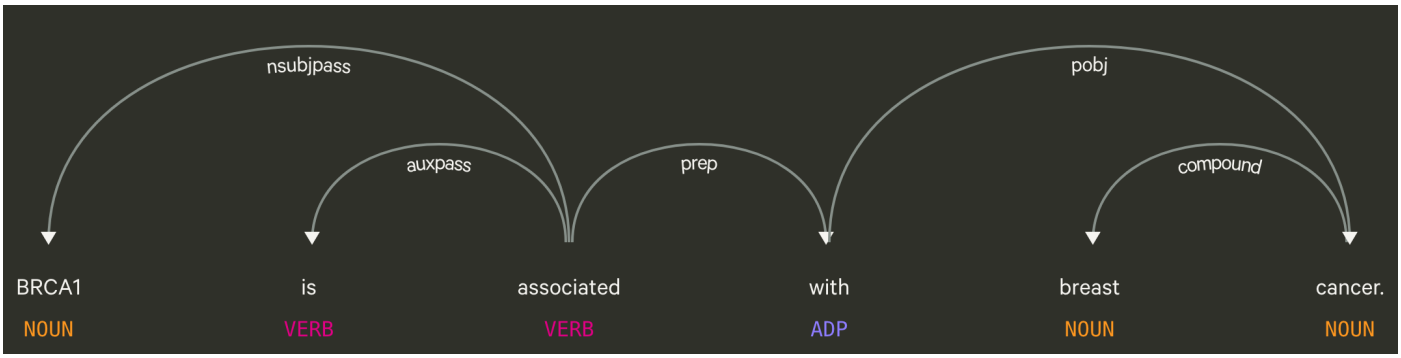

Figure 2: A visualization of a dependency parse tree using the following sentence as in example: “BRCA1 is associated with breast cancer” [64]. For these type of trees the root begins at the main verb of a sentence. Each arrows depicts the dependency shared between two words. For example, the dependency between BRCA1 and associated is nsubjpass, which stands for passive nominal subject. This means that BRCA1 is the subject of the sentneces and it is being referred to by the word associated.

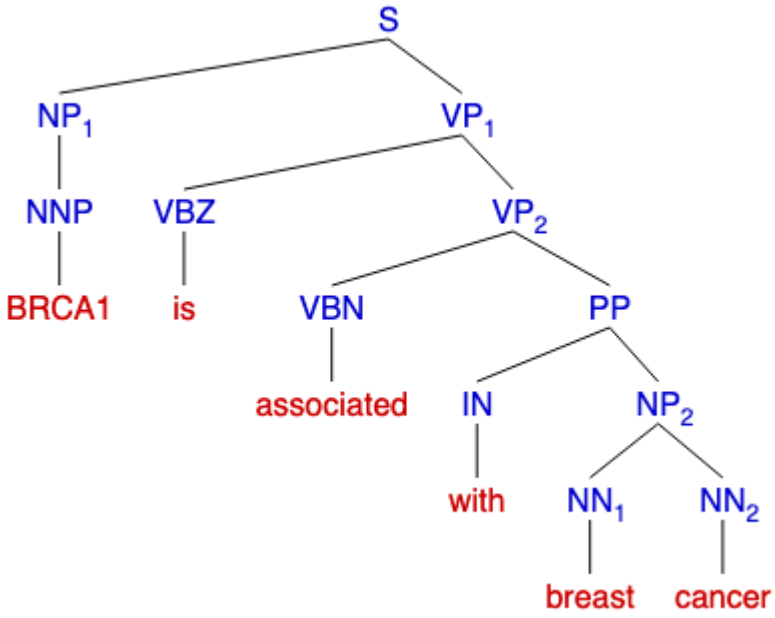

Figure 3: A visualization of a constituency parse tree using the following sentence: "BRCA1 is associated with breast cancer" [65]. This type of tree has the root beginning at the start of the sentence. Each word is grouped into subphrases depending on the part of speech tags of a word. For example, the word "associated" is a past particple verb (VBN) that belongs to the verb phrase (VP) subgroup.

Extracting Relationships Without Labels

Unsupervised methods of extraction involve drawing inferences from data without the use of labels. These methods involve some form of clustering or statistical calculations. In this section we discuss methods that use unsupervised learning to detect relationship asserting sentences from text.

An unsupervised method to extract relationships exploits the fact that two entities can appear together in text. This kind of event is called co-occurrence and studies that use this phenomenon can be found in table 1. Two databases DISEASES [12] and STRING [66] were populated using a co-occurrence scoring method on PubMed abstracts. Both databases used the same scoring method that measured the frequency of co-mention pairs within individual sentences as well as the abstracts themselves. This method assumes independence between each individual occurrence. Under this assumption mention pairs that occur more than expected were presumed to indicate the presence of an association or interaction. This approach was able to identify 543,405 disease gene associations [12] and 792,730 high confidence protein protein interactions [66], but is limited to only using PubMed abstracts.

Full text articles are able to drastically amplify text mining power to detect relationships [67,68]. Westergaard et al. used a co-occurrence approach, similar to DISEASES [12] and STRING [66], to mine full articles for protein-protein interactions and other protein related information [67]. The authors discovered that full text provided better prediction power than using abstracts alone. This improvement suggests that future text mining approaches should consider using full text to increase detection power.

Unsupervised methods have been focused on treating multiple biomedical relationships as multiple isolated problems. These methods repeatedly use the same model for each biomedical relationship type. An alternative to this persepctive is to capture all different relationship types at once. Clustering is an approach that accomplish this concept of simultaneous extraction. Percha et al. used a biclustering algorithm on generated dependency parse trees to group PubMed abstract sentences [69]. Each cluster was manually curated to determine which relationship they represented. This approach captured 4,451,661 dependency paths for 36 different groups [69]. Despite the success, this approach suffered from technical issues such as dependency tree parsing errors. This type of error

resulted in sentences not being grouped by the clustering algorithm [69]. Future clustering approaches should consider simplifying sentences to prevent this type of issue.

Overall unsupervised methods provide a means to rapidly find relationship asserting sentences without the need of annotated text. Approaches in this category range from using co-occurrence scores to clustering sentences. These methods provide a generalizable framework that can be used on large repositories of text. Future methods can improve detection power by considering the use of methods that simplify sentences and use datasets that include full text articles.

Table 1: Table of approaches that mainly use a form of co-occurrence.

Study	Relationship of Interest
[70]	Protein-Protein Interactions, Disease-Gene and Tissue-Gene Associations
[71]	Drug Disease Treatments
[72]	Drug, Gene and Disease interactions
[67]	Protein-Protein Interactions
[12]	Disease-Gene associations
[73]	Protein-Protein Interactions
[74]	Genotype-Phenotype Relationships

Supervised Relationship Extraction

Supervised extraction uses labeled relationships to learn text patterns that correspond to positively labeled relationships instead of negative ones. Most of these approaches have flourished due to pre-labelled publicly available datasets (Table 2). These datasets were constructed by curators for shared open tasks [75,76] or as a means to provide the scientific community with a gold standard [76,77,78]. Approaches that use these available datasets range from using support vector machines (SVMs) with custom kernels to deep learning with algorithms that can construct their own features. In the rest of this section we discuss approaches that use supervised methods to detect relationship-asserting sentences.

Extracting relationships in a supervised setting can involve mapping textual input onto a high dimensional space. Support vector machines are a type of classifier that can accomplish this task with a mapping function called a kernel [52,79]. These kernels take information such as a sentence's dependency tree [50,51], part of speech tags [52] or even word counts [79] and map them onto a dense feature space. Within this space, the methods learn a hyperplane that separates sentences in the positive class (mentions a relationship) from the negative class (does not mention a relationship). Kernels can be manually constructed or selected to cater to the relationship being extracted [51,52,79,79]. Determining the correct kernel requires expert knowledge to be successful and is a nontrivial task depending on the relationship. In addition to single kernel methods, a recent study used an ensemble of SVMs to extract disease-gene associations [80]. The ensemble outperformed notable disease-gene association extractors [63,81] in terms of precision, recall and F1 score. Overall, SVMs have been shown to be beneficial in terms of relationship mining; however, major focus have shifted to utilizing deep learning techniques to extract relationships as these approaches can perform non-linear mappings of high dimensional data.

Deep learning is an increasingly popular class of techniques that can construct their own features within a high dimensional space [82,83]. These methods use different forms of neural networks, such as recurrent or convolutional neural networks, to perform classification.

Recurrent neural networks (RNN) are designed for sequential analysis that consist of using a repeatedly updating hidden state to make predictions. An example of a recurrent neural network is a long short term memory (LSTM) network [84]. Cocos et al [85] used a LSTM to extract drug side effects from de-identified twitter posts, while Yadav et al. [86] used an LSTM to extract protein-protein interactions. Other works have used LSTMs to perform relationship extraction [85,87,88,89,90]. Despite the success of these networks, training can be difficult as these networks are highly susceptible to vanishing and exploding gradients [91,92]. One solution to this problem is to clip the gradients while the neural network trains [93]. Besides the gradient problem, these approaches peak in performance when the dataset reaches at least a tens of thousand of data points [94].

Convolutional neural networks (CNNs), which are widely applied for image analysis, use multiple kernel filters to capture small subsets of an overall image [83]. In the context of text mining an image is replaced with words within a sentence mapped to dense vectors (i.e., word embeddings) [95,96]. Peng et al. [97] used a CNN to extract sentences that mentioned protein-protein interactions and Zhou et al. [98] used a CNN to extract chemical-disease relations. Other approaches have used CNNs and variants of CNNs to extract relationship-asserting sentences [100,101,99]. Just like RNNs, these networks perform well when millions of labeled examples are present [94]. Future approaches that use CNNs or RNNs should consider solutions to obtaining these large quantities of data through means such as weak supervision [102], semi-supervised learning [103] or using pre-trained networks via transfer learning [104,105].

Semi-supervised learning [103] and weak supervision [102] are techniques that can construct large datasets for machine learning classifiers. Semi-supervised learning consists of combining labeled data with unlabeled data to extract relationships. For example, one study used a variational auto encoder with a LSTM network to extract protein-protein interactions from pubmed abstracts and full text [106]. This is an elegant solution to handle the small dataset problem, but requires labeled data to start. The dependency on labeled data makes finding under-studied relationships difficult as one would need to find or construct examples of the missing relationships in the beginning.

Weak or distant supervision takes a different approach that uses noisy or even erroneous labels to train classifiers [102,107,108,109]. Under this paradigm sentences are labeled based on their mention pair being present (positive) or absent (negative) in a database. Once these labels are obtained a machine learning classifier can now be trained to predict sentences [102]. For example, Thomas et al. [110] used distant supervision to train a support vector machine to extract sentences mentioning protein-protein interactions (ppi). Their SVM model achieved comparable performance against a baseline model; however, the noise generated via distant supervision was difficult to eradicate [110]. A number of efforts have focused on combining distant supervision with other types of labeling strategies to reduce the negative impacts of noisy knowledge bases [111,112,113]. Nicholson et al. [101] found that, in some circumstances, these strategies and rules can be reused across different types of biomedical edges to learn a heterogeneous knowledge graph if those edges describe similar physical concepts. This remains an active area of investigation with numerous associated challenges and opportunities. Overall, semi-supervised learning and weak supervision provide promising results in terms of relation extraction and future approaches should consider using those paradigms to train machine learning classifiers.

Table 2: A set of publicly available datasets for supervised text mining.

Dataset	Type of Sentences
AIMed [41]	PPI
BioInfer [114]	PPI
LLL [115]	PPI
IEPA [116]	PPI

Dataset	Type of Sentences
HPRD5 [77]	PPI
EU-ADR [40]	DaG
BeFree [81]	DaG
CoMAGC [78]	DaG
CRAFT [117]	DaG
Biocreative V CDR [76]	Compound induces Disease
Biocreative IV ChemProt [75]	CbG

Applying Knowledge Graphs to Biomedical Challenges

1. Mention that these graphs can be used for discovery
2. Mention representation learning (aka representing a graph as dense vectors for nodes and/or edges)
3.

Unifying Techniques

Mapping high dimensional data into a low dimensional space has greatly improved modeling performance in fields such as natural language processing [95,96] and image analysis [118]. The success of these approaches provides rationale for projecting knowledge graphs into a low dimensional space as well [119]. Techniques that perform this projection often require information on how nodes are connected with one another [120,121,122,123], while other approaches can work directly with the edges themselves [124]. We group methods for producing low-dimensional representations of knowledge graphs into the following three categories: matrix factorization, translational methods, and deep learning (Figure 4).

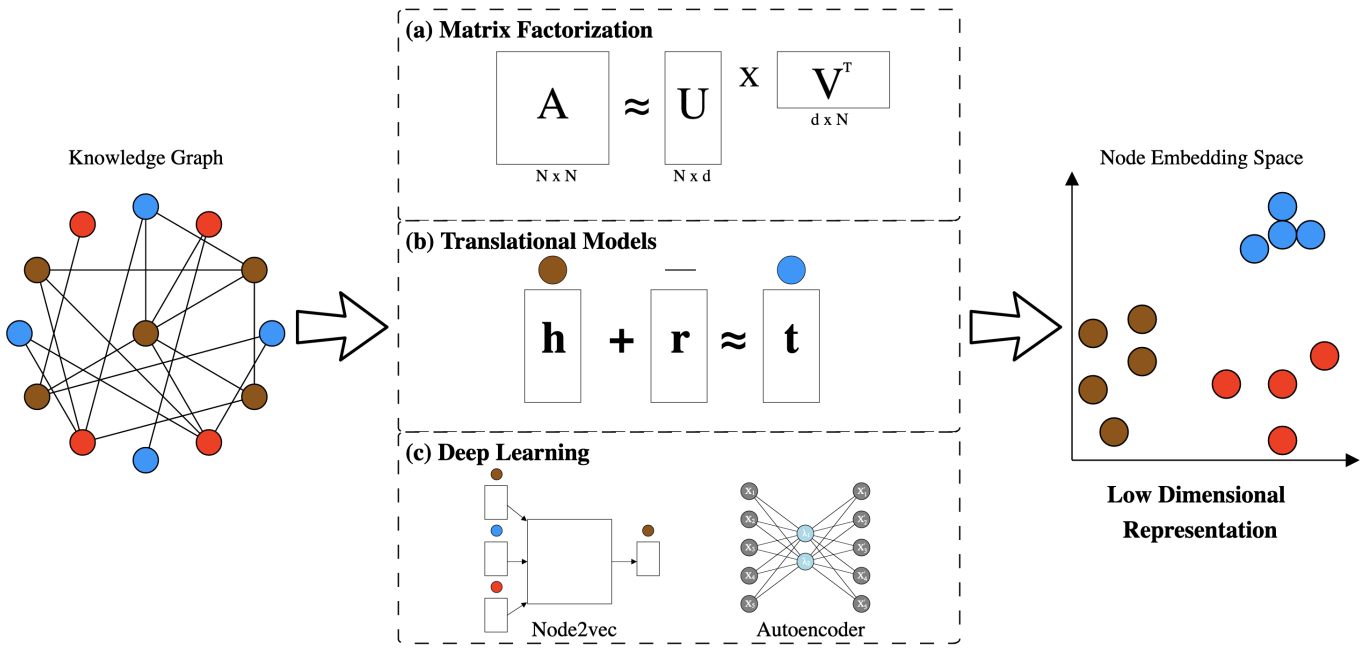

Figure 4: Pipeline for embedding knowledge graphs into a low dimensional space. Starting with a knowledge graph, embeddings can be generated using one of the following options: Matrix Factorization (a), Translational Models (b) or Deep Learning (c). The output of this pipeline is an embedding space that clusters similar node types together.

Matrix Factorization

Matrix factorization is a technique that uses linear algebra to map high dimensional data into a low dimensional space. This projection is accomplished by decomposing a matrix into a set of small rectangular matrices (Figure 4 (a)). Notable methods for matrix decomposition include Isomap [125], Laplacian eigenmaps [126] and Principal Component Analysis (PCA) [127]/Singular Vector Decomposition (SVD) [128]. These methods were designed to be used on many different types of data; however, we discuss their use in the context of projecting knowledge graphs into a low dimensional space.

SVD [128] is an algorithm that uses matrix factorization to represent knowledge graphs in a low dimensional space. The input for this algorithm is an adjacency matrix (A), which is a square matrix where rows and columns represent nodes and each entry represents the presence of an edge between two nodes. This adjacency matrix (A) gets decomposed into three parts: a square matrix Σ and a set of two small rectangular matrices U and V^T. This values within Σ are called singular values, which akin to eigenvalues [128]. Each row in U and each column in V^T represents nodes projected onto a low dimensional space [127,128]. In practice U is usually used to represent nodes in a knowledge graph, but V^T can also be used [128,129]. Typically, SVD appears in recommendation systems via collaborative filtering [130]; however, this technique can also be used as a standalone baseline to compare to other approaches [131].

Laplacian eigenmaps assume there is low dimensional structure in a high dimensional space [126]. This algorithm preserves this structure while projecting data into a low dimensional space. Typically, the first step of this algorithm is to construct a figurative knowledge graph where nodes represent datapoints and edges are constructed based on similarity of two datapoints; however, in this context, the knowledge graph has already been constructed. The next step in this algorithm is to obtain both an adjacency matrix (A) and a degree matrix (D) from the knowledge graph. A degree matrix is a diagonal matrix where each entry represents the number of edges connected to a node. The adjacency and degree matrices are converted into a laplacian matrix (L), which is a matrix that shares the same properties as the adjacency matrix. The laplacian matrix is generated by subtracting the adjacency matrix from the degree matrix ($L = D - A$) and, once constructed, the algorithm uses linear algebra to calculate eigenvalues and eigenvectors from the matrix ($Lx = \lambda Dx$). The generated eigenvectors represent the knowledge graph's nodes projected onto a low dimensional space [126]. A number of approaches have used variants of this algorithm to perform their own node projection [120,121,132]. Typically, eigenmaps work well when knowledge graphs have a sparse number of edges between nodes but struggle when presented with denser networks [131,132,133]. A future direction is to adapt these methods to scale to knowledge graphs that have a large number of edges.

Matrix factorization is a powerful technique that uses a matrices such as an adjacency matrix as input. Common approaches involve using SVD, Laplacian eigenmaps or variants of the two to perform embeddings. Despite reported success, the dependence on matrices like an adjacency matrix creates an issue of scalability as matrices of large networks would take too much memory for a regular computer to handle. Furthermore, these methods treat all edge types the same, but a possible extension for future approaches that use matrix factorization would be to incorporate node and edge types as sources of input.

Translational Distance Models

Translational distance models treat edges in a knowledge graph as linear transformations. As an example, one such algorithm, TransE [134], treats every node-edge pair as a triplet with head nodes represented as $\mathbf{h}$, edges represented as $\mathbf{r}$, and tail nodes represented as $\mathbf{t}$. These representations are combined into an equation that mimics the iconic word vectors translations ($\mathbf{king} - \mathbf{man} + \mathbf{woman} \approx \mathbf{queen}$) from the Word2vec model [96]. The equation is shown as follows: $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$. Starting at the head node ($\mathbf{h}$), add the edge vector ($\mathbf{r}$) and the result should be the tail node ($\mathbf{t}$). TransE optimizes embeddings for $\mathbf{h}$, $\mathbf{r}$, $\mathbf{t}$, while guaranteeing the global equation (

$\mathbf{h} + \mathbf{r} \approx \mathbf{t}$) is satisfied [134]. A caveat to the TransE approach is that it the training steps force relationships to have a one to one mapping, which may not be appropriate for all types of relationships.

Wang et al. [135] attempted to resolve the one to one mapping issue by developing the TransH model. TransH treats relations as hyperplanes rather than a regular vector and projects the head ($\mathbf{h}$) and tail ($\mathbf{t}$) nodes onto the hyperplane. Following this projection, a distance vector ($\mathbf{d}_r$) is calculated between the projected head and tail nodes. Finally, each vector is optimized while preserving the global equation ($\mathbf{h} + \mathbf{d}_r \approx \mathbf{t}$) [135]. Other approaches [136,137] have built off of the TransE and TransH models. In the future, it may be beneficial for these models is to incorporate other types of information such as edge confidence scores, textual information, or edge type information when optimizing these embeddings.

Deep Learning

Deep learning is a paradigm that uses multiple non-linear transformations to map high dimensional data into a low dimensional space. Many techniques that use deep learning for knowledge graphs are based on word2vec [95,96], a set of approaches that are widely used for natural language processing. The goal of word2vec is to project words into a low dimensional space that preserves their semantic meaning. Strategies for training word2vec models use one of two neural network architectures: skip-gram and continuous bag of words (CBOW). Both models are feed-forward neural networks, but CBOW models are trained to predict a word given it's context while skip-gram models are trained to predict the context given a word [95,96]. Once training has finished, words are now associated with dense vectors that downstream models, such as feed forward networks or recurrent networks, can use for input.

Deepwalk [138] is an early method designed to project a knowledge graph into a low dimensional space. The first step of this method is to perform a random walk along a knowledge graph. During the random walk, every generated sequence of nodes is recorded and treated like a sentence in word2vec [95,96]. After every node has been processed, a skip-gram model is trained to predict the context of each node thereby projecting a knowledge graph into a low dimensional space [138]. A limitation of this method is that the random walk cannot be controlled, so every node has an equal chance to be reached. Grover and Leskovec [139] demonstrated that this limitation can hurt performance when classifying edges between nodes and developed node2vec as a result. Node2vec [139] operates the in the same fashion as deepwalk; however, this algorithm specifies a parameter that lets the random walk be biased when traversing nodes. A caveat to both deepwalk and node2vec is that both algorithms ignore information such as edge type and node type. Various approaches have evolved to fix this limitation by incorporating node, edge and even path types when projecting nodes into a low dimensional space [140,141,142,143]. These approaches primarily capture a network's local structure. An emerging area of work is to develop approaches that capture both the local and global structure of a network when projecting knowledge graphs into a low dimensional space.

Some deep learning approaches use an adjacency matrix as input [95,96] instead of using the word2vec framing. Algorithms such as auto-encoders can also generate network embeddings [144,145,146]. Autoencoders [147,148] are neural networks that map input such as an adjacency matrices into a low dimensional space and then learns how to construct this space by reconstructing the same input. The generated low dimensional space captures the node connectivity structure of the knowledge graph and every node is mapped onto this space [144,145,146]. Despite the high potential of this approach, this method relies on an adjacency matrix for input. If a knowledge graph asymptotically increases in size, these approaches could run into scalability issues as discovered by Khosla et al. [149]. Plus, Khosla et al.[149] discovered that approaches akin to node2vec outperformed algorithms using autoencoders when undergoing link prediction and node classification. Overall, the performance of these models largely depends upon the structure of nodes

and edges within a knowledge graph [149]. Future approaches should consider creating hybrid models that use both node2vec and autoencoders to construct complementary low dimensional representations of knowledge graphs.

Unifying Applications

Knowledge graphs have been used in many biomedical applications ranging from identifying protein functions [150] to prioritizing cancer genes [151] to recommending safer drugs to patients [152,153]. In this section we discuss how knowledge graphs are being applied in biomedical settings. We put particular emphasis on an emerging set of techniques: those that project knowledge graphs into a low dimensional space.

Disease and Gene Interactions

1. Mention disease gene prioritization
2. Mention Disease gene associations

Protein Protein Interactions

1. Mention predicting genes interacting genes

Drug Interactions

1. Talk about drug side effects
2. Drug repurposing
3. Drug-Disease Interations

Clinical applications

1. Can mention EHR use and other related applications
2. Mention Tiffany's work on private data embeddings

Conclusion

1. Summarize discussed positives and pitfalls
2. Leave some open ended questions yet to be explored
3. Will come into play as I write this review paper

References

1. **Node Classification in Social Networks**
Smriti Bhagat, Graham Cormode, S. Muthukrishnan
Social Network Data Analytics (2011) https://doi.org/fjj48w
DOI: 10.1007/978-1-4419-8462-3_5

2. **Network Embedding Based Recommendation Method in Social Networks**
Yufei Wen, Lei Guo, Zhumin Chen, Jun Ma
Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18 (2018)
https://doi.org/gf6rtt
DOI: 10.1145/3184558.3186904

3. **Open Question Answering with Weakly Supervised Embedding Models**
Antoine Bordes, Jason Weston, Nicolas Usunier
arXiv (2014-04-16) https://arxiv.org/abs/1404.4326v1

4. **Neural Network-based Question Answering over Knowledge Graphs on Word and Character Level**
Denis Lukovnikov, Asja Fischer, Jens Lehmann, Sören Auer
Proceedings of the 26th International Conference on World Wide Web - WWW '17 (2017)
https://doi.org/gfv8hp
DOI: 10.1145/3038912.3052675

5. **Towards integrative gene prioritization in Alzheimer's disease.**
Jang H Lee, Graciela H Gonzalez
Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing (2011)
https://www.ncbi.nlm.nih.gov/pubmed/21121028
DOI: 10.1142/9789814335058_0002 · PMID: 21121028

6. **PhenoGeneRanker: A Tool for Gene Prioritization Using Complete Multiplex Heterogeneous Networks**
Cagatay Dursun, Naoki Shimoyama, Mary Shimoyama, Michael Schläppi, Serdar Bozdag
Cold Spring Harbor Laboratory (2019-05-27) https://doi.org/gf6rtr
DOI: 10.1101/651000

7. **Biological Random Walks: Integrating heterogeneous data in disease gene prioritization**
Michele Gentili, Leonardo Martini, Manuela Petti, Lorenzo Farina, Luca Becchetti
2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) (2019-07) https://doi.org/gf6rts
DOI: 10.1109/cibcb.2019.8791472

8. **Semantic Disease Gene Embeddings (SmuDGE): phenotype-based disease gene prioritization without phenotypes**
Mona Alshahrani, Robert Hoehndorf
Bioinformatics (2018-09-01) https://doi.org/gd9k8n
DOI: 10.1093/bioinformatics/bty559 · PMID: 30423077 · PMCID: PMC6129260

9. **Systematic integration of biomedical knowledge prioritizes drugs for repurposing**
Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, Sergio E Baranzini

eLife (2017-09-22) https://doi.org/cdfk
DOI: 10.7554/elife.26726 · PMID: 28936969 · PMCID: PMC5640425

10. **Assessing Drug Target Association Using Semantic Linked Data**
Bin Chen, Ying Ding, David J. Wild
PLoS Computational Biology (2012-07-05) https://doi.org/rn6
DOI: 10.1371/journal.pcbi.1002574 · PMID: 22859915 · PMCID: PMC3390390

11. **Towards a definition of knowledge graphs**
Lisa Ehrlinger, Wolfram Wöß
SEMANTiCS (2016)

12. **DISEASES: Text mining and data integration of disease–gene associations**
Sune Pletscher-Frankild, Albert Pallejà, Kalliopi Tsafou, Janos X. Binder, Lars Juhl Jensen
Methods (2015-03) https://doi.org/f3mn6s
DOI: 10.1016/j.ymeth.2014.11.020 · PMID: 25484339

13. **DrugBank 5.0: a major update to the DrugBank database for 2018**
David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, … Michael Wilson
Nucleic Acids Research (2017-11-08) https://doi.org/gcwtzk
DOI: 10.1093/nar/gkx1037 · PMID: 29126136 · PMCID: PMC5753335

14. **A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information**
Yunan Luo, Xinbin Zhao, Jingtian Zhou, Jinglin Yang, Yanqing Zhang, Wenhua Kuang, Jian Peng, Ligong Chen, Jianyang Zeng
Nature Communications (2017-09-18) https://doi.org/gbxwrc
DOI: 10.1038/s41467-017-00680-8 · PMID: 28924171 · PMCID: PMC5603535

15. **Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks**
Hui Liu, Yinglong Song, Jihong Guan, Libo Luo, Ziheng Zhuang
BMC Bioinformatics (2016-12) https://doi.org/gf6v27
DOI: 10.1186/s12859-016-1336-7 · PMID: 28155639 · PMCID: PMC5259862

16. **Finding disease similarity based on implicit semantic similarity**
Sachin Mathur, Deendayal Dinakarpandian
Journal of Biomedical Informatics (2012-04) https://doi.org/b7b3tw
DOI: 10.1016/j.jbi.2011.11.017 · PMID: 22166490

17. **KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences**
Patrick Ernst, Amy Siu, Gerhard Weikum
BMC Bioinformatics (2015-05-14) https://doi.org/gb8w8d
DOI: 10.1186/s12859-015-0549-5 · PMID: 25971816 · PMCID: PMC4448285

18. **Constructing biomedical domain-specific knowledge graph with minimum supervision**
Jianbo Yuan, Zhiwei Jin, Han Guo, Hongxia Jin, Xianchao Zhang, Tristram Smith, Jiebo Luo
Knowledge and Information Systems (2019-03-23) https://doi.org/gf6v26
DOI: 10.1007/s10115-019-01351-4

19. **Feature assisted stacked attentive shortest dependency path based Bi-LSTM model for protein–protein interaction**
Shweta Yadav, Asif Ekbal, Sriparna Saha, Ankit Kumar, Pushpak Bhattacharyya
Knowledge-Based Systems (2019-02) https://doi.org/gf4788
DOI: 10.1016/j.knosys.2018.11.020

20. **Biological Databases- Integration of Life Science Data**
Nishant Toomula, Arun Kumar, Sathish Kumar D, Vijaya Shanti Bheemidi
Journal of Computer Science & Systems Biology (2012) https://doi.org/gf8qcb
DOI: 10.4172/jcsb.1000081

21. **COSMIC: somatic cancer genetics at high-resolution**
Simon A. Forbes, David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G. Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, … Peter J. Campbell
Nucleic Acids Research (2016-11-28) https://doi.org/f9v865
DOI: 10.1093/nar/gkw1121 · PMID: 27899578 · PMCID: PMC5210583

22. **COSMIC: the Catalogue Of Somatic Mutations In Cancer**
John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, … Simon A Forbes
Nucleic Acids Research (2018-10-29) https://doi.org/gf9hxg
DOI: 10.1093/nar/gky1015 · PMID: 30371878 · PMCID: PMC6323903

23. **Recurated protein interaction datasets**
Lukasz Salwinski, Luana Licata, Andrew Winter, David Thorneycroft, Jyoti Khadake, Arnaud Ceol, Andrew Chatr Aryamontri, Rose Oughtred, Michael Livstone, Lorrie Boucher, … Henning Hermjakob
Nature Methods (2009-12) https://doi.org/fgvkmf
DOI: 10.1038/nmeth1209-860 · PMID: 19935838

24. **Literature-curated protein interaction datasets**
Michael E Cusick, Haiyuan Yu, Alex Smolyar, Kavitha Venkatesan, Anne-Ruxandra Carvunis, Nicolas Simonis, Jean-François Rual, Heather Borick, Pascal Braun, Matija Dreze, … Marc Vidal
Nature Methods (2008-12-30) https://doi.org/d4j62p
DOI: 10.1038/nmeth.1284 · PMID: 19116613 · PMCID: PMC2683745

25. **Curation accuracy of model organism databases**
I. M. Keseler, M. Skrzypek, D. Weerasinghe, A. Y. Chen, C. Fulcher, G.-W. Li, K. C. Lemmer, K. M. Mladinich, E. D. Chow, G. Sherlock, P. D. Karp
Database (2014-06-12) https://doi.org/gf63jz
DOI: 10.1093/database/bau058 · PMID: 24923819 · PMCID: PMC4207230

26. **OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders**
Joanna S. Amberger, Carol A. Bocchini, François Schiettecatte, Alan F. Scott, Ada Hamosh
Nucleic Acids Research (2014-11-26) https://doi.org/gf8qb6
DOI: 10.1093/nar/gku1205 · PMID: 25428349 · PMCID: PMC4383985

27. **Textpresso Central: a customizable platform for searching, text mining, viewing, and curating biomedical literature**
H.-M. Müller, K. M. Van Auken, Y. Li, P. W. Sternberg
BMC Bioinformatics (2018-03-09) https://doi.org/gf7rbz
DOI: 10.1186/s12859-018-2103-8 · PMID: 29523070 · PMCID: PMC5845379

28. **Text mining and expert curation to develop a database on psychiatric diseases and their genes**
Alba Gutiérrez-Sacristán, Àlex Bravo, Marta Portero-Tresserra, Olga Valverde, Antonio Armario, M. C. Blanco-Gandía, Adriana Farré, Lierni Fernández-Ibarrondo, Francina Fonseca, Jesús Giraldo, … Laura I. Furlong
Database (2017-01-01) https://doi.org/gf8qb5
DOI: 10.1093/database/bax043 · PMID: 29220439 · PMCID: PMC5502359

29. **Manual curation is not sufficient for annotation of genomic databases**
William A. Baumgartner Jr, K. Bretonnel Cohen, Lynne M. Fox, George Acquaah-Mensah, Lawrence Hunter
Bioinformatics (2007-07-01) https://doi.org/dtck86
DOI: 10.1093/bioinformatics/btm229 · PMID: 17646325 · PMCID: PMC2516305

30. **The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index**
Peder Olesen Larsen, Markus von Ins
Scientometrics (2010-03-10) https://doi.org/c4hb8r
DOI: 10.1007/s11192-010-0202-z · PMID: 20700371 · PMCID: PMC2909426

31. **Semi-automatic semantic annotation of PubMed queries: A study on quality, efficiency, satisfaction**
Aurélie Névéol, Rezarta Islamaj Doğan, Zhiyong Lu
Journal of Biomedical Informatics (2011-04) https://doi.org/bq34sj
DOI: 10.1016/j.jbi.2010.11.001 · PMID: 21094696 · PMCID: PMC3063330

32. **Assisting manual literature curation for protein-protein interactions using BioQRator**
D. Kwon, S. Kim, S.-Y. Shin, A. Chatr-aryamontri, W. J. Wilbur
Database (2014-07-22) https://doi.org/gf7hm3
DOI: 10.1093/database/bau067 · PMID: 25052701 · PMCID: PMC4105708

33. **Argo: an integrative, interactive, text mining-based workbench supporting curation**
R. Rak, A. Rowley, W. Black, S. Ananiadou
Database (2012-03-20) https://doi.org/h5d
DOI: 10.1093/database/bas010 · PMID: 22434844 · PMCID: PMC3308166

34. **CurEx**
Michael Loster, Felix Naumann, Jan Ehmueller, Benjamin Feldmann
Proceedings of the 27th ACM International Conference on Information and Knowledge Management - CIKM '18 (2018) https://doi.org/gf8qb8
DOI: 10.1145/3269206.3269229

35. **Re-curation and rational enrichment of knowledge graphs in Biological Expression Language**
Charles Tapley Hoyt, Daniel Domingo-Fernández, Rana Aldisi, Lingling Xu, Kristian Kolpeja, Sandra Spalek, Esther Wollert, John Bachman, Benjamin M Gyori, Patrick Greene, Martin Hofmann-Apitius
Database (2019-01-01) https://doi.org/gf7hm4
DOI: 10.1093/database/baz068 · PMID: 31225582 · PMCID: PMC6587072

36. **LocText: relation extraction of protein localizations to assist database curation**
Juan Miguel Cejuela, Shrikant Vinchurkar, Tatyana Goldberg, Madhukar Sollepura Prabhu Shankar, Ashish Baghudana, Aleksandar Bojchevski, Carsten Uhlig, André Ofner, Pandu Raharja-Liu, Lars Juhl Jensen, Burkhard Rost

BMC Bioinformatics (2018-01-17) https://doi.org/gf8qb9
DOI: 10.1186/s12859-018-2021-9 · PMID: 29343218 · PMCID: PMC5773052

37. **Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements**
Todd Lingren, Louise Deleger, Katalin Molnar, Haijun Zhai, Jareen Meinzen-Derr, Megan Kaiser, Laura Stoutenborough, Qi Li, Imre Solti
Journal of the American Medical Informatics Association (2014-05) https://doi.org/f5zggh
DOI: 10.1136/amiajnl-2013-001837 · PMID: 24001514 · PMCID: PMC3994857

38. **iSimp in BioC standard format: enhancing the interoperability of a sentence simplification system**
Y. Peng, C. O. Tudor, M. Torii, C. H. Wu, K. Vijay-Shanker
Database (2014-05-21) https://doi.org/gf9hxf
DOI: 10.1093/database/bau038 · PMID: 24850848 · PMCID: PMC4028706

39. **BioSimplify: an open source sentence simplification engine to improve recall in automatic biomedical information extraction.**
Siddhartha Jonnalagadda, Graciela Gonzalez
AMIA ... Annual Symposium proceedings. AMIA Symposium (2010-11-13)
https://www.ncbi.nlm.nih.gov/pubmed/21346999
PMID: 21346999 · PMCID: PMC3041388

40. **The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships**
Erik M. van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A. Kors, Laura I. Furlong
Journal of Biomedical Informatics (2012-10) https://doi.org/f36vn6
DOI: 10.1016/j.jbi.2012.04.004 · PMID: 22554700

41. **Comparative experiments on learning information extractors for proteins and their interactions**
Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun K. Ramani, Yuk Wah Wong
Artificial Intelligence in Medicine (2005-02) https://doi.org/dhztpn
DOI: 10.1016/j.artmed.2004.07.016 · PMID: 15811782

42. **A Unified Active Learning Framework for Biomedical Relation Extraction**
Hong-Tao Zhang, Min-Lie Huang, Xiao-Yan Zhu
Journal of Computer Science and Technology (2012-11) https://doi.org/gf8qb4
DOI: 10.1007/s11390-012-1306-0

43. **Entrez Gene: gene-centered information at NCBI**
D. Maglott, J. Ostell, K. D. Pruitt, T. Tatusova
Nucleic Acids Research (2010-11-28) https://doi.org/fsjcqz
DOI: 10.1093/nar/gkq1237 · PMID: 21115458 · PMCID: PMC3013746

44. **UniProt: a worldwide hub of protein knowledge***Nucleic Acids Research* (2018-11-05)
https://doi.org/gfwqck
DOI: 10.1093/nar/gky1049 · PMID: 30395287 · PMCID: PMC6323992

45. **Pharmacogenomics Knowledge for Personalized Medicine**
M Whirl-Carrillo, EM McDonagh, JM Hebert, L Gong, K Sangkuhl, CF Thorn, RB Altman, TE Klein

Clinical Pharmacology & Therapeutics (2012-10) https://doi.org/gdnfzr
DOI: 10.1038/clpt.2012.96 · PMID: 22992668 · PMCID: PMC3660037

46. **The BioGRID interaction database: 2013 update**
Andrew Chatr-aryamontri, Bobby-Joe Breitkreutz, Sven Heinicke, Lorrie Boucher, Andrew Winter, Chris Stark, Julie Nixon, Lindsay Ramage, Nadine Kolas, Lara O'Donnell, … Mike Tyers
Nucleic Acids Research (2012-11-30) https://doi.org/f4jmz4
DOI: 10.1093/nar/gks1158 · PMID: 23203989 · PMCID: PMC3531226

47. **The Comparative Toxicogenomics Database: update 2019**
Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Roy McMorran, Jolene Wiegers, Thomas C Wiegers, Carolyn J Mattingly
Nucleic Acids Research (2018-09-24) https://doi.org/gf8qb7
DOI: 10.1093/nar/gky868 · PMID: 30247620 · PMCID: PMC6323936

48. **CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database**
Baofeng Jia, Amogelang R. Raphenya, Brian Alcock, Nicholas Waglechner, Peiyao Guo, Kara K. Tsang, Briony A. Lago, Biren M. Dave, Sheldon Pereira, Arjun N. Sharma, … Andrew G. McArthur
Nucleic Acids Research (2016-10-26) https://doi.org/f9wbjs
DOI: 10.1093/nar/gkw1004 · PMID: 27789705 · PMCID: PMC5210516

49. **OMIM.org: leveraging knowledge across phenotype-gene relationships.**
Joanna S Amberger, Carol A Bocchini, Alan F Scott, Ada Hamosh
Nucleic acids research (2019-01-08) https://www.ncbi.nlm.nih.gov/pubmed/30445645
DOI: 10.1093/nar/gky1151 · PMID: 30445645 · PMCID: PMC6323937

50. **LPTK: a linguistic pattern-aware dependency tree kernel approach for the BioCreative VI CHEMPROT task**
Neha Warikoo, Yung-Chun Chang, Wen-Lian Hsu
Database (2018-01-01) https://doi.org/gfhjr6
DOI: 10.1093/database/bay108 · PMID: 30346607 · PMCID: PMC6196310

51. **DTMiner: identification of potential disease targets through biomedical literature mining**
Dong Xu, Meizhuo Zhang, Yanping Xie, Fan Wang, Ming Chen, Kenny Q. Zhu, Jia Wei
Bioinformatics (2016-08-09) https://doi.org/f9nw36
DOI: 10.1093/bioinformatics/btw503 · PMID: 27506226 · PMCID: PMC5181534

52. **Exploiting graph kernels for high performance biomedical relation extraction**
Nagesh C. Panyam, Karin Verspoor, Trevor Cohn, Kotagiri Ramamohanarao
Journal of Biomedical Semantics (2018-01-30) https://doi.org/gf49nn
DOI: 10.1186/s13326-017-0168-3 · PMID: 29382397 · PMCID: PMC5791373

53. **iSimp in BioC standard format: enhancing the interoperability of a sentence simplification system.**
Yifan Peng, Catalina O Tudor, Manabu Torii, Cathy H Wu, K Vijay-Shanker
Database : the journal of biological databases and curation (2014-05-21)
https://www.ncbi.nlm.nih.gov/pubmed/24850848
DOI: 10.1093/database/bau038 · PMID: 24850848 · PMCID: PMC4028706

54. **BELMiner: adapting a rule-based relation extraction system to extract biological expression language statements from bio-medical literature evidence sentences**
K. E. Ravikumar, Majid Rastegar-Mojarad, Hongfang Liu

Database (2017-01-01) https://doi.org/gf7rbx
DOI: 10.1093/database/baw156 · PMID: 28365720 · PMCID: PMC5467463

55. **A generalizable NLP framework for fast development of pattern-based biomedical relation extraction systems**
Yifan Peng, Manabu Torii, Cathy H Wu, K Vijay-Shanker
BMC Bioinformatics (2014-08-23) https://doi.org/f6rndz
DOI: 10.1186/1471-2105-15-285 · PMID: 25149151 · PMCID: PMC4262219

56. **Construction of phosphorylation interaction networks by text mining of full-length articles using the eFIP system**
Catalina O. Tudor, Karen E. Ross, Gang Li, K. Vijay-Shanker, Cathy H. Wu, Cecilia N. Arighi
Database (2015-01-01) https://doi.org/gf8fpt
DOI: 10.1093/database/bav020 · PMID: 25833953 · PMCID: PMC4381107

57. **miRTex: A Text Mining System for miRNA-Gene Relation Extraction**
Gang Li, Karen E. Ross, Cecilia N. Arighi, Yifan Peng, Cathy H. Wu, K. Vijay-Shanker
PLOS Computational Biology (2015-09-25) https://doi.org/f75mwb
DOI: 10.1371/journal.pcbi.1004391 · PMID: 26407127 · PMCID: PMC4583433

58. **LimTox: a web tool for applied text mining of adverse event and toxicity associations of compounds, drugs and genes**
Andres Cañada, Salvador Capella-Gutierrez, Obdulia Rabal, Julen Oyarzabal, Alfonso Valencia, Martin Krallinger
Nucleic Acids Research (2017-05-22) https://doi.org/gf479h
DOI: 10.1093/nar/gkx462 · PMID: 28531339 · PMCID: PMC5570141

59. **DiMeX: A Text Mining System for Mutation-Disease Association Extraction**
A. S. M. Ashique Mahmood, Tsung-Jung Wu, Raja Mazumder, K. Vijay-Shanker
PLOS ONE (2016-04-13) https://doi.org/f8xktj
DOI: 10.1371/journal.pone.0152725 · PMID: 27073839 · PMCID: PMC4830514

60. **Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors**
F. Horn, A. L. Lau, F. E. Cohen
Bioinformatics (2004-01-22) https://doi.org/d7cjgj
DOI: 10.1093/bioinformatics/btg449 · PMID: 14990452

61. **Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing**
Rong Xu, QuanQiu Wang
BMC Bioinformatics (2013-06-06) https://doi.org/gb8v3k
DOI: 10.1186/1471-2105-14-181 · PMID: 23742147 · PMCID: PMC3702428

62. **RLIMS-P 2.0: A Generalizable Rule-Based Information Extraction System for Literature Mining of Protein Phosphorylation Information**
Manabu Torii, Cecilia N. Arighi, Gang Li, Qinghua Wang, Cathy H. Wu, K. Vijay-Shanker
IEEE/ACM Transactions on Computational Biology and Bioinformatics (2015-01-01) https://doi.org/gf8fpv
DOI: 10.1109/tcbb.2014.2372765 · PMID: 26357075 · PMCID: PMC4568560

63. **PKDE4J: Entity and relation extraction for public knowledge discovery**
Min Song, Won Chul Kim, Dahee Lee, Go Eun Heo, Keun Young Kang

Journal of Biomedical Informatics (2015-10) https://doi.org/f7v7jj
DOI: 10.1016/j.jbi.2015.08.008 · PMID: 26277115

64. **Spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing**
Matthew Honnibal, Ines Montani
To appear (2017)

65. **PhpSyntaxTree tool**
A Eisenbach, M Eisenbach
(2006)

66. **STRING v9.1: protein-protein interaction networks, with increased coverage and integration**
Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian von Mering, Lars J. Jensen
Nucleic Acids Research (2012-11-29) https://doi.org/gf5kcd
DOI: 10.1093/nar/gks1094 · PMID: 23203871 · PMCID: PMC3531103

67. **A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts**
David Westergaard, Hans-Henrik Stærfeldt, Christian Tønsberg, Lars Juhl Jensen, Søren Brunak
PLOS Computational Biology (2018-02-15) https://doi.org/gcx747
DOI: 10.1371/journal.pcbi.1005962 · PMID: 29447159 · PMCID: PMC5831415

68. **STITCH 4: integration of protein–chemical interactions with user data**
Michael Kuhn, Damian Szklarczyk, Sune Pletscher-Frankild, Thomas H. Blicher, Christian von Mering, Lars J. Jensen, Peer Bork
Nucleic Acids Research (2013-11-28) https://doi.org/f5shb4
DOI: 10.1093/nar/gkt1207 · PMID: 24293645 · PMCID: PMC3964996

69. **A global network of biomedical relationships derived from text**
Bethany Percha, Russ B Altman
Bioinformatics (2018-02-27) https://doi.org/gc3ndk
DOI: 10.1093/bioinformatics/bty114 · PMID: 29490008 · PMCID: PMC6061699

70. **CoCoScore: context-aware co-occurrence scoring for text mining applications using distant supervision**
Alexander Junge, Lars Juhl Jensen
Bioinformatics (2019-06-14) https://doi.org/gf4789
DOI: 10.1093/bioinformatics/btz490 · PMID: 31199464 · PMCID: PMC6956794

71. **A new method for prioritizing drug repositioning candidates extracted by literature-based discovery**
Majid Rastegar-Mojarad, Ravikumar Komandur Elayavilli, Dingcheng Li, Rashmi Prasad, Hongfang Liu
2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2015-11)
https://doi.org/gf479j
DOI: 10.1109/bibm.2015.7359766

72. **Literature Mining for the Discovery of Hidden Connections between Drugs, Genes and Diseases**
Raoul Frijters, Marianne van Vugt, Ruben Smeets, René van Schaik, Jacob de Vlieg, Wynand Alkema
PLoS Computational Biology (2010-09-23) https://doi.org/bhrw7x
DOI: 10.1371/journal.pcbi.1000943 · PMID: 20885778 · PMCID: PMC2944780

73. **STRING v10: protein–protein interaction networks, integrated over the tree of life**
Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P. Tsafou, ... Christian von Mering
Nucleic Acids Research (2014-10-28) https://doi.org/f64rfn
DOI: 10.1093/nar/gku1003 · PMID: 25352553 · PMCID: PMC4383874

74. **Text Mining Genotype-Phenotype Relationships from Biomedical Literature for Database Curation and Precision Medicine**
Ayush Singhal, Michael Simmons, Zhiyong Lu
PLOS Computational Biology (2016-11-30) https://doi.org/f9gz4b
DOI: 10.1371/journal.pcbi.1005017 · PMID: 27902695 · PMCID: PMC5130168

75. **Overview of the biocreative vi chemical-protein interaction track**
Martin Krallinger, Obdulia Rabal, Saber A Akhondi, others
Proceedings of the sixth biocreative challenge evaluation workshop (2017)
https://www.semanticscholar.org/paper/Overview-of-the-BioCreative-VI-chemical-protein-Krallinger-Rabal/eed781f498b563df5a9e8a241c67d63dd1d92ad5

76. **BioCreative V CDR task corpus: a resource for chemical disease relation extraction**
Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, Zhiyong Lu
Database (2016) https://doi.org/gf5hfw
DOI: 10.1093/database/baw068 · PMID: 27161011 · PMCID: PMC4860626

77. **RelEx–Relation extraction using dependency parse trees**
K. Fundel, R. Kuffner, R. Zimmer
Bioinformatics (2006-12-01) https://doi.org/cz7q4d
DOI: 10.1093/bioinformatics/btl616 · PMID: 17142812

78. **CoMAGC: a corpus with multi-faceted annotations of gene-cancer relations**
Hee-Jin Lee, Sang-Hyung Shim, Mi-Ryoung Song, Hyunju Lee, Jong C Park
BMC Bioinformatics (2013) https://doi.org/gb8v5s
DOI: 10.1186/1471-2105-14-323 · PMID: 24225062 · PMCID: PMC3833657

79. **Text Mining for Protein Docking**
Varsha D. Badal, Petras J. Kundrotas, Ilya A. Vakser
PLOS Computational Biology (2015-12-09) https://doi.org/gcvj3b
DOI: 10.1371/journal.pcbi.1004630 · PMID: 26650466 · PMCID: PMC4674139

80. **Automatic extraction of gene-disease associations from literature using joint ensemble learning**
Balu Bhasuran, Jeyakumar Natarajan
PLOS ONE (2018-07-26) https://doi.org/gdx63f
DOI: 10.1371/journal.pone.0200699 · PMID: 30048465 · PMCID: PMC6061985

81. **Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research**
Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, Laura I Furlong
BMC Bioinformatics (2015-02-21) https://doi.org/f7kn8s
DOI: 10.1186/s12859-015-0472-9 · PMID: 25886734 · PMCID: PMC4466840

82. **Deep learning**
Ian Goodfellow, Yoshua Bengio, Aaron Courville
The MIT Press (2016)
ISBN: 0262035618, 9780262035613

83. **Deep learning**
Yann LeCun, Yoshua Bengio, Geoffrey Hinton
Nature (2015-05) https://doi.org/bmqp
DOI: 10.1038/nature14539 · PMID: 26017442

84. **Long Short-Term Memory**
Sepp Hochreiter, Jürgen Schmidhuber
Neural Computation (1997-11) https://doi.org/bxd65w
DOI: 10.1162/neco.1997.9.8.1735 · PMID: 9377276

85. **Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts**
Anne Cocos, Alexander G Fiks, Aaron J Masino
Journal of the American Medical Informatics Association (2017-02-22) https://doi.org/gbp9nj
DOI: 10.1093/jamia/ocw180 · PMID: 28339747

86. **Semantic Relations in Compound Nouns: Perspectives from Inter-Annotator Agreement**
Yadav Prabha, Jezek Elisabetta, Bouillon Pierrette, Callahan Tiffany J., Bada Michael, Hunter Lawrence E., Cohen K. Bretonnel
Studies in Health Technology and Informatics (2017) https://doi.org/ggmk8t
DOI: 10.3233/978-1-61499-830-3-644

87. **Cross-Sentence N-ary Relation Extraction with Graph LSTMs**
Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, Wen-tau Yih
arXiv (2017-08-12) https://arxiv.org/abs/1708.03743v1

88. **Drug drug interaction extraction from biomedical literature using syntax convolutional neural network**
Zhehuan Zhao, Zhihao Yang, Ling Luo, Hongfei Lin, Jian Wang
Bioinformatics (2016-07-27) https://doi.org/f9nsq7
DOI: 10.1093/bioinformatics/btw486 · PMID: 27466626 · PMCID: PMC5181565

89. **N-ary Relation Extraction using Graph State LSTM**
Linfeng Song, Yue Zhang, Zhiguo Wang, Daniel Gildea
arXiv (2018-08-28) https://arxiv.org/abs/1808.09101v1

90. **A neural joint model for entity and relation extraction from biomedical text**
Fei Li, Meishan Zhang, Guohong Fu, Donghong Ji
BMC Bioinformatics (2017-03-31) https://doi.org/gcgnx2
DOI: 10.1186/s12859-017-1609-9 · PMID: 28359255 · PMCID: PMC5374588

91. **The problem of learning long-term dependencies in recurrent networks**
Y. Bengio, P. Frasconi, P. Simard
IEEE International Conference on Neural Networks https://doi.org/d7zs24
DOI: 10.1109/icnn.1993.298725

92. **Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network**

Alex Sherstinsky
arXiv (2018-08-09) https://arxiv.org/abs/1808.03314v5
DOI: 10.1016/j.physd.2019.132306

93. **On the difficulty of training Recurrent Neural Networks**
Razvan Pascanu, Tomas Mikolov, Yoshua Bengio
arXiv (2012-11-21) https://arxiv.org/abs/1211.5063v2

94. **Revisiting Unreasonable Effectiveness of Data in Deep Learning Era**
Chen Sun, Abhinav Shrivastava, Saurabh Singh, Abhinav Gupta
arXiv (2017-07-10) https://arxiv.org/abs/1707.02968v2

95. **Efficient Estimation of Word Representations in Vector Space**
Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean
arXiv (2013-01-16) https://arxiv.org/abs/1301.3781v3

96. **Distributed Representations of Words and Phrases and their Compositionality**
Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean
arXiv (2013-10-16) https://arxiv.org/abs/1310.4546v1

97. **Deep learning for extracting protein-protein interactions from biomedical literature**
Yifan Peng, Zhiyong Lu
arXiv (2017-06-05) https://arxiv.org/abs/1706.01556v2

98. **Knowledge-guided convolutional networks for chemical-disease relation extraction**
Huiwei Zhou, Chengkun Lang, Zhuang Liu, Shixian Ning, Yingyu Lin, Lei Du
BMC Bioinformatics (2019-05-21) https://doi.org/gf45zn
DOI: 10.1186/s12859-019-2873-7 · PMID: 31113357 · PMCID: PMC6528333

99. **Extraction of protein–protein interactions (PPIs) from the literature by deep convolutional neural networks with various feature embeddings**
Sung-Pil Choi
Journal of Information Science (2016-11-01) https://doi.org/gcv8bn
DOI: 10.1177/0165551516673485

100. **Extracting chemical–protein relations with ensembles of SVM and deep learning models**
Yifan Peng, Anthony Rios, Ramakanth Kavuluru, Zhiyong Lu
Database (2018-01-01) https://doi.org/gf479f
DOI: 10.1093/database/bay073 · PMID: 30020437 · PMCID: PMC6051439

101. **Expanding a Database-derived Biomedical Knowledge Graph via Multi-relation Extraction from Biomedical Abstracts**
David N. Nicholson, Daniel S. Himmelstein, Casey S. Greene
Cold Spring Harbor Laboratory (2019-08-08) https://doi.org/gf6qxh
DOI: 10.1101/730085

102. **Distant supervision for relation extraction without labeled data**
Mike Mintz, Steven Bills, Rion Snow, Dan Jurafsky
Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - ACL-IJCNLP '09 (2009) https://doi.org/fg9q43
DOI: 10.3115/1690219.1690287

103. **Introduction to Semi-Supervised Learning**
Xiaojin Zhu, Andrew B. Goldberg
Synthesis Lectures on Artificial Intelligence and Machine Learning (2009-01) https://doi.org/bq7pm2
DOI: 10.2200/s00196ed1v01y200906aim006

104. **A Survey on Transfer Learning**
Sinno Jialin Pan, Qiang Yang
IEEE Transactions on Knowledge and Data Engineering (2010-10) https://doi.org/bc4vws
DOI: 10.1109/tkde.2009.191

105. **A survey of transfer learning**
Karl Weiss, Taghi M. Khoshgoftaar, DingDing Wang
Journal of Big Data (2016-05-28) https://doi.org/gfkr2w
DOI: 10.1186/s40537-016-0043-6

106. **Exploring Semi-supervised Variational Autoencoders for Biomedical Relation Extraction**
Yijia Zhang, Zhiyong Lu
arXiv (2019-01-18) https://arxiv.org/abs/1901.06103v1

107. **Large-scale extraction of gene interactions from full-text literature using DeepDive**
Emily K. Mallory, Ce Zhang, Christopher Ré, Russ B. Altman
Bioinformatics (2015-09-03) https://doi.org/gb5g7b
DOI: 10.1093/bioinformatics/btv476 · PMID: 26338771 · PMCID: PMC4681986

108. **Snorkel**
Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, Christopher Ré
Proceedings of the VLDB Endowment (2017-11-01) https://doi.org/ch44
DOI: 10.14778/3157794.3157797 · PMID: 29770249 · PMCID: PMC5951191

109. **Snorkel MeTaL**
Alex Ratner, Braden Hancock, Jared Dunnmon, Roger Goldman, Christopher Ré
Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning - DEEM'18 (2018) https://doi.org/gf3xk7
DOI: 10.1145/3209889.3209898 · PMID: 30931438 · PMCID: PMC6436830

110. **Learning protein protein interaction extraction using distant supervision**
Philippe Thomas, Illés Solt, Roman Klinger, Ulf Leser
(2011-01)

111. **Robust Distant Supervision Relation Extraction via Deep Reinforcement Learning**
Pengda Qin, Weiran Xu, William Yang Wang
arXiv (2018-05-24) https://arxiv.org/abs/1805.09927v1

112. **DSGAN: Generative Adversarial Training for Distant Supervision Relation Extraction**
Pengda Qin, Weiran Xu, William Yang Wang
arXiv (2018-05-24) https://arxiv.org/abs/1805.09929v1

113. **Noise Reduction Methods for Distantly Supervised Biomedical Relation Extraction**
Gang Li, Cathy Wu, K. Vijay-Shanker
BioNLP 2017 (2017) https://doi.org/ggmk8s
DOI: 10.18653/v1/w17-2323

114. **BioInfer: a corpus for information extraction in the biomedical domain**
Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, Tapio Salakoski
BMC Bioinformatics (2007-02-09) https://doi.org/b7bhhc
DOI: 10.1186/1471-2105-8-50 · PMID: 17291334 · PMCID: PMC1808065

115. **Learning language in logic - genic interaction extraction challenge**
C. Nédellec
Proceedings of the learning language in logic 2005 workshop at the international conference on machine learning (2005)

116. **Mining medline: Abstracts, sentences, or phrases?**
Jing Ding, Daniel Berleant, Dan Nettleton, Eve Syrkin Wurtele
Pacific symposium on biocomputing (2002) http://helix-web.stanford.edu/psb02/ding.pdf

117. **Concept annotation in the CRAFT corpus**
Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner Jr, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, Lawrence E Hunter
BMC Bioinformatics (2012-07-09) https://doi.org/gb8vdr
DOI: 10.1186/1471-2105-13-161 · PMID: 22776079 · PMCID: PMC3476437

118. **Deep Residual Learning for Image Recognition**
Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun
arXiv (2015-12-10) https://arxiv.org/abs/1512.03385v1

119. **Representation Learning on Graphs: Methods and Applications**
William L. Hamilton, Rex Ying, Jure Leskovec
arXiv (2017-09-17) https://arxiv.org/abs/1709.05584v3

120. **Signed laplacian embedding for supervised dimension reduction**
Chen Gong, Dacheng Tao, Jie Yang, Keren Fu
Proceedings of the twenty-eighth aaai conference on artificial intelligence (2014)
http://dl.acm.org/citation.cfm?id=2892753.2892809

121. **A Semi-NMF-PCA Unified Framework for Data Clustering**
Kais Allab, Lazhar Labiod, Mohamed Nadif
IEEE Transactions on Knowledge and Data Engineering (2017-01-01) https://doi.org/f9hm9g
DOI: 10.1109/tkde.2016.2606098

122. **Partially supervised graph embedding for positive unlabelled feature selection**
Yufei Han, Yun Shen
Proceedings of the twenty-fifth international joint conference on artificial intelligence (2016)
http://dl.acm.org/citation.cfm?id=3060832.3060837
ISBN: 978-1-57735-770-4

123. **GraRep**
Shaosheng Cao, Wei Lu, Qiongkai Xu
Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15 (2015) https://doi.org/gf8rgf
DOI: 10.1145/2806416.2806512

124. **Improved Knowledge Base Completion by Path-Augmented TransR Model**
Wenhao Huang, Ge Li, Zhi Jin
arXiv (2016-10-06) https://arxiv.org/abs/1610.04073v1

125. **A Global Geometric Framework for Nonlinear Dimensionality Reduction**
J. B. Tenenbaum
Science (2000-12-22) https://doi.org/cz8wgk
DOI: 10.1126/science.290.5500.2319 · PMID: 11125149

126. **Laplacian Eigenmaps for Dimensionality Reduction and Data Representation**
Mikhail Belkin, Partha Niyogi
Neural Computation (2003-06) https://doi.org/bbr9cw
DOI: 10.1162/089976603321780317

127. **Principal component analysis**
Svante Wold, Kim Esbensen, Paul Geladi
Chemometrics and Intelligent Laboratory Systems (1987-08) https://doi.org/bm8dnf
DOI: 10.1016/0169-7439(87)80084-9

128. **The approximation of one matrix by another of lower rank**
Carl Eckart, Gale Young
Psychometrika (1936-09) https://doi.org/c2frtd
DOI: 10.1007/bf02288367

129. **Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec**
Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, Jie Tang
Proceedings of the eleventh acm international conference on web search and data mining (2018)
https://doi.org/10.1145/3159652.3159706
DOI: 10.1145/3159652.3159706 · ISBN: 9781450355810

130. **A Survey of Collaborative Filtering Techniques**
Xiaoyuan Su, Taghi M. Khoshgoftaar
Advances in Artificial Intelligence (2009) https://doi.org/fk9jjg
DOI: 10.1155/2009/421425

131. **Graph Embedding on Biomedical Networks: Methods, Applications, and Evaluations**
Xiang Yue, Zhen Wang, Jingong Huang, Srinivasan Parthasarathy, Soheil Moosavinasab, Yungui Huang, Simon M. Lin, Wen Zhang, Ping Zhang, Huan Sun
arXiv (2019-06-12) https://arxiv.org/abs/1906.05017v3
DOI: 10.1093/bioinformatics/btz718

132. **GLEE: Geometric Laplacian Eigenmap Embedding**
Leo Torres, Kevin S Chan, Tina Eliassi-Rad
arXiv (2019-05-23) https://arxiv.org/abs/1905.09763v2

133. **Vicus: Exploiting local structures to improve network-based analysis of biological data**
Bo Wang, Lin Huang, Yuke Zhu, Anshul Kundaje, Serafim Batzoglou, Anna Goldenberg
PLOS Computational Biology (2017-10-12) https://doi.org/gb368p
DOI: 10.1371/journal.pcbi.1005621 · PMID: 29023470 · PMCID: PMC5638230

134. **Translating embeddings for modeling multi-relational data**
Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, Oksana Yakhnenko
NIPS (2013)

135. **Knowledge graph embedding by translating on hyperplanes**
Zhen Wang, Jianwen Zhang, Jianlin Feng, Zheng Chen

Proceedings of the twenty-eighth aaai conference on artificial intelligence (2014)
http://dl.acm.org/citation.cfm?id=2893873.2894046

136. **Learning entity and relation embeddings for knowledge graph completion**
Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, Xuan Zhu
Proceedings of the twenty-ninth aaai conference on artificial intelligence (2015)
http://dl.acm.org/citation.cfm?id=2886521.2886624
ISBN: 0-262-51129-0

137. **PrTransH: Embedding Probabilistic Medical Knowledge from Real World EMR Data**
Linfeng Li, Peng Wang, Yao Wang, Jinpeng Jiang, Buzhou Tang, Jun Yan, Shenghui Wang, Yuting Liu
arXiv (2019-09-02) https://arxiv.org/abs/1909.00672v1

138. **DeepWalk: Online Learning of Social Representations**
Bryan Perozzi, Rami Al-Rfou, Steven Skiena
arXiv (2014-03-26) https://arxiv.org/abs/1403.6652v2
DOI: 10.1145/2623330.2623732

139. **node2vec: Scalable Feature Learning for Networks**
Aditya Grover, Jure Leskovec
arXiv (2016-07-03) https://arxiv.org/abs/1607.00653v1

140. **struc2vec: Learning Node Representations from Structural Identity**
Leonardo F. R. Ribeiro, Pedro H. P. Savarese, Daniel R. Figueiredo
arXiv (2017-04-11) https://arxiv.org/abs/1704.03165v3
DOI: 10.1145/3097983.3098061

141. **metapath2vec**
Yuxiao Dong, Nitesh V. Chawla, Ananthram Swami
Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17 (2017) https://doi.org/gfsqzn
DOI: 10.1145/3097983.3098036

142. **edge2vec: Representation learning using edge semantics for biomedical knowledge discovery**
Zheng Gao, Gang Fu, Chunping Ouyang, Satoshi Tsutsui, Xiaozhong Liu, Jeremy Yang, Christopher Gessner, Brian Foote, David Wild, Qi Yu, Ying Ding
arXiv (2018-09-07) https://arxiv.org/abs/1809.02269v3

143. **Learning Graph Embeddings from WordNet-based Similarity Measures**
Andrey Kutuzov, Mohammad Dorgham, Oleksiy Oliynyk, Chris Biemann, Alexander Panchenko
arXiv (2018-08-16) https://arxiv.org/abs/1808.05611v4

144. **Learning to Make Predictions on Graphs with Autoencoders**
Phi Vu Tran
arXiv (2018-02-23) https://arxiv.org/abs/1802.08352v2
DOI: 10.1109/dsaa.2018.00034

145. **Variational Graph Auto-Encoders**
Thomas N. Kipf, Max Welling
arXiv (2016-11-21) https://arxiv.org/abs/1611.07308v1

146. **Adversarially Regularized Graph Autoencoder for Graph Embedding**
Shirui Pan, Ruiqi Hu, Guodong Long, Jing Jiang, Lina Yao, Chengqi Zhang
arXiv (2018-02-13) https://arxiv.org/abs/1802.04407v2

147. **Deep Learning in Neural Networks: An Overview**
Juergen Schmidhuber
arXiv (2014-04-30) https://arxiv.org/abs/1404.7828v4
DOI: 10.1016/j.neunet.2014.09.003

148. **Autoencoders, unsupervised learning and deep architectures**
Pierre Baldi
Proceedings of the 2011 international conference on unsupervised and transfer learning workshop - volume 27 (2011)

149. **A Comparative Study for Unsupervised Network Representation Learning**
Megha Khosla, Vinay Setty, Avishek Anand
arXiv (2019-03-19) https://arxiv.org/abs/1903.07902v5
DOI: 10.1109/tkde.2019.2951398

150. **Neural networks for link prediction in realistic biomedical graphs: a multi-dimensional evaluation of graph embedding-based approaches**
Gamal Crichton, Yufan Guo, Sampo Pyysalo, Anna Korhonen
BMC Bioinformatics (2018-05-21) https://doi.org/ggkm7q
DOI: 10.1186/s12859-018-2163-9 · PMID: 29783926 · PMCID: PMC5963080

151. **Network-based integration of multi-omics data for prioritizing cancer genes**
Christos Dimitrakopoulos, Sravanth Kumar Hindupur, Luca Häfliger, Jonas Behr, Hesam Montazeri, Michael N Hall, Niko Beerenwinkel
Bioinformatics (2018-03-14) https://doi.org/gc6953
DOI: 10.1093/bioinformatics/bty148 · PMID: 29547932 · PMCID: PMC6041755

152. **Safe Medicine Recommendation via Medical Knowledge Graph Embedding**
Meng Wang, Mengyue Liu, Jun Liu, Sen Wang, Guodong Long, Buyue Qian
arXiv (2017-10-16) https://arxiv.org/abs/1710.05980v2

153. **GAMENet: Graph Augmented MEmory Networks for Recommending Medication Combination**
Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, Jimeng Sun
Proceedings of the AAAI Conference on Artificial Intelligence (2019-07-17) https://doi.org/ggkm7r
DOI: 10.1609/aaai.v33i01.33011126