ERLKG: Entity Representation Learning and Knowledge Graph based association analysis of COVID-19 through mining of unstructured biomedical corpora

Sayantan Basu [1] **Sinchani Chakraborty** [2] **Atif Hassan** [3]
Sana Siddique [4] **Ashish Anand** [5]
[1,5] Indian Institute of Technology Guwahati
[2,3] Indian Institute of Technology Kharagpur
[4] Eras Lucknow Medical College and Hospital
[1] sayantan18@iitg.ac.in [2]sinchanichakraborty@gmail.com

Abstract

We introduce a generic, human-out-of-the-loop pipeline, ERLKG, to perform rapid association analysis of any biomedical entity with other existing entities from a corpora of the same domain. Our pipeline consists of a Knowledge Graph (KG) created from the Open Source CORD-19 dataset by fully automating the procedure of information extraction using SciBERT. The best latent entity representations are then found by benchnmarking different KG embedding techniques on the task of link prediction using a Graph Convolution Network Auto Encoder (GCN-AE). We demonstrate the utility of ERLKG with respect to COVID-19 through multiple qualitative evaluations. Due to the lack of a gold standard, we propose a relatively large intrinsic evaluation dataset for COVID-19 and use it for validating the top two performing KG embedding techniques. We find TransD to be the best performing KG embedding technique with Pearson and Spearman correlation scores of 0.4348 and 0.4570 respectively. We demonstrate that a considerable number of ERLKG's top protein, chemical and disease predictions are currently in consideration for COVID-19 related research.

1 Introduction

COVID-19 is a global epidemic with a considerable fatality rate and a high transmission rate, affecting millions of people world-wide since its outbreak.[1] The search for treatments and possible cures for the novel Coronavirus (Wang et al., 2020b) has led to an exponential increase in scientific publications, but the challenge lies in effectively processing, integrating and leveraging related sources of information.

Rapid and effective utilization of literature during times of pandemic such as COVID-19 is of utmost importance in combating the disease. In this paper, we introduce a fully automated generic pipeline consisting of an Information Extraction (IE) system followed by Knowledge Graph construction. The IE module uses SciBERT (Beltagy et al., 2019) for performing Named Entity Recognition (NER) and Relationship Extraction (RE). The entire entity extraction procedure is fully automated and no human expertise is used. The major goal is to ensure rapid access of relevant data through a structured representation of free text articles. Following this, we focus on the task of association analysis of essential biomedical entities, namely, proteins, diseases and, chemicals. Such entities are well explored in existing literature and an analysis of their relatedness to COVID-19 is provided by leveraging the CORD-19 Open Research Dataset (Wang et al., 2020a). This can assist the physicians to accelerate knowledge discovery and provide support for clinical decision making. The dataset and related resources of this paper are made public[2].

Due to a lack of gold standard information, we perform extensive qualitative evaluations in order to show that our system does not suffer from redundancy or bias. These evaluations include performance on a link prediction task and intrinsic evaluation. For the former, KG embeddings along with graph adjacency matrix are fed to a GCN-AE (Kipf and Welling, 2016) model to perform link prediction. Average Precision (AP) and ROC scores were used to benchmark different KG embeddings on the generated knowledge graph. For the intrinsic evaluation, we propose a new dataset that has been developed with the help of three physicians and benchmark our embeddings against it. Finally,

[1] https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200811-covid-19-sitrep-204.pdf?sfvrsn=1f4383dd_2

[2] https://github.com/sayantanbasu05/ERKLG

Proceedings of the First Workshop on Scholarly Document Processing, pages 127–137
Online, November 19, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

based on cosine similarity score, the best representation was used to predict top chemicals, proteins and diseases related to COVID-19. The contributions of our approach are as follows :

1. We propose a fully automated, human-out-of-the-loop, end-to-end generic pipeline for rapidly determining association of any biomedical entity of interest with other existing well explored entities.

2. We benchmark multiple KG embedding techniques on the task of link prediction and demonstrate that simple embedding methods provide comparable performance on straightforward structured KGs.

3. We introduce two human gold-standard entity lists, COV19_25 and COV19_729. The former consists of expert ratings for 25 entities predicted by ERLKG while the latter consists of expert ratings for 729 entities sampled from the CORD-19 dataset. The ratings are based on every entity's relatedness with respect to COVID-19.

2 Related Work

We mostly focus on recent works centered around the CORD-19 dataset by discussing about the techniques used for IE and KG generation.

2.1 Entity and Relation Extraction

Most of the recent NLP systems use pretrained language models on unannotated text like ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), and XLNet (Yang et al., 2019). In the biomedical and clinical domains, BERT based architectures pretrained with domain-specific unlabelled text have been used for IE (Lee et al., 2020; Alsentzer et al., 2019). The CORD-19 dataset, curated for the COVID-19 pandemic, integrates related scientific articles for various information retrieval tasks (Roberts et al., 2020). Multiple NLP applications have been developed around CORD-19 like Question Answering (Das et al., 2020), Summarization (Park, 2020), NER (Wang et al., 2020c), etc.

2.2 Knowledge Graph

KGs were immensely used in different fields like Life Science (Chen et al., 2009), Decision Support System (Russell and Norvig, 2010) etc. Using the CORD-19 dataset and many other textual sources, KGs have been built and used for performing different tasks that aid in knowledge discovery. Chen et al. (2020) performs NER using BioBERT on CORD-19 and PubMed Dataset (Dernoncourt and Lee, 2017) while developing a Coronavirus KG from PubMed KG based on two different methods, namely, cosine similarity and co-occurence frequency to predict plausible drugs. Wang et al. (2020b) construct a KG termed as COVID-KG, by extracting multimodal knowledge from existing scientific literature and ontology followed by a QA system, built on top of this information, with an aim to answer questions related to drug repurposing. Comparatively smaller KGs have been constructed for COVID-19 like (Domingo-Fernández et al., 2020) which covers 145 articles consisting of 3945 nodes and 9484 relations covering 10 entity types. Previously built KGs have also been employed for COVID-19 drug discovery (Richardson et al., 2020). However, the scope of the network built by the last two methods is limited owing to the smaller dataset size. Also, to learn node representations and leverage the structural information of the graph, various techniques are used for Knowledge Graph embeddings. Rossi et al. (2020) conducts extensive survey on 16 KG embedding techniques to perform a comparative analysis. They form a taxonomy of the embedding methods, grouping various methods to tensor decomposition models like DisMult (Yang et al., 2015), Geometric models like TransE (Bordes et al., 2013), TransD (Ji et al., 2015), ComplEx (Trouillon et al., 2016) and Rotate (Sun et al., 2019) and Deep Learning models like ConVE (Dettmers et al., 2018) and CapsE (Nguyen et al., 2019). Shifting from textual source to construct a KG, Ray et al. (2020) uses biological interaction networks like drug-protein and protein-protein networks to predict repurposable drugs for SARS-CoV-2 through link prediction while employing Variational Graph AutoEncoders with features from Node2Vec (Grover and Leskovec, 2016) for entity representation.

3 Dataset

3.1 CORD-19

The CORD-19 corpus (Wang et al., 2020a) was published by Allen AI in association with White House and other organizations. It was made publicly available on the Kaggle [3] platform as a part

[3] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

of an open research challenge. The data, containing scholarly articles, is collected from sources like PubMed Central (PMC), PubMed, the World Health Organization's COVID-19 Database, and various preprint servers like bioRxiv, medRxiv and arXiv.

CORD-19 corpus (2020-05-12) contains a pool of 1,38,000 scholarly articles with 69,000 full-text articles related to COVID-19, SARS-CoV-2, etc. Each paper is associated with bibliographic metadata such as Title, Author etc, as well as unique identifiers such as a DOI, PubMed Central ID etc. Various sub-tasks have been identified for effective information retrieval, however, it lacks task oriented ground truth data. We merge all the metadata with corresponding full text papers and retain the title, abstract and full text from the corpus.

3.2 Datasets for Fine-tuning SciBERT

For NER, we consider the following three datasets, namely, JNLPBA (Collier and Kim, 2004) corpus which consists of 5 distinct tags: *Protein, DNA, RNA, Cell line and Cell type*, the CHEMDNER (Krallinger et al., 2015) corpus which consists of : *Abbreviation, Family, Formula, Identifiers, Multiple, Systematic and Trivial*, the NCBI Disease Corpus (Dogan et al., 2014) which is used to identify only disease mentions.

For RE, the following datasets are used, namely, CHEMPROT (Kringelum et al., 2016) which consists of 13 different relationship types based on identified positive associations according to : *Inhibitor, Substrate, Indirect-Down regulator, Indirect-Up regulator, Activator, Antagonist, Product-Of, Agonist, Down regulator, Up regulator, Agonist-Activator, Agonist-Inhibitor and Substrator-Product-Of* and BC5CDR (Li et al., 2016) which captures binary relations predicting positive or negative interaction for chemical-induced-disease pairs.

4 ERLKG

In this section we discuss about the entire pipeline and its various components. Figure 1 depicts the pipeline which consists of the following modules : Preprocessing, Named Entity Recognition (NER), Relation Extraction (RE) and Knowledge Graph (KG) construction. The rest part of the Figure 1 depicts the evaluation strategies adopted for a reliable association analysis of various chemical, protein and drug entities from CORD-19 corpus with respect to COVID-19.

4.1 Preprocessing

Each abstract or full text was split into sentences using NLTK (Loper and Bird, 2002) sentence tokenizer and the sentences, in turn, were tokenized using the Spacy (v2.0.10) tokenizer[4]. Following this we removed all the non-functional tokens and attached POS tags to the remaining tokens.

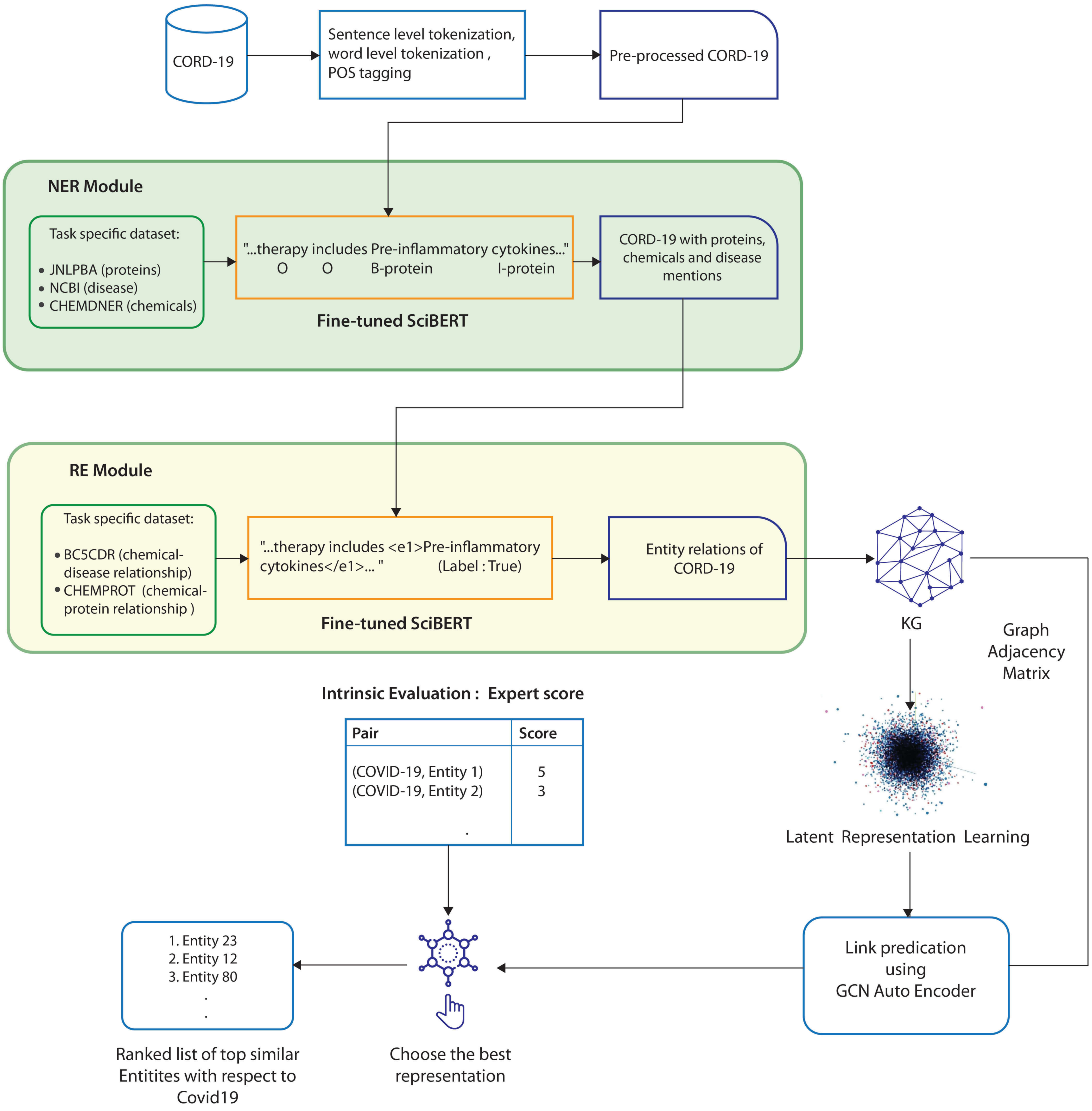
4.2 Named Entity Recognition

Named Entity Recognition (NER) is the task of identifying domain-specific proper nouns in a sentence. In order to gain meaningful insights about the major classes of biomedical entities present in the dataset, it was necessary to tag the entities using an NER module by fine-tuning on various biomedical datasets. Since the CORD-19 dataset is a collection of scientific articles, we use SciBERT for NER extraction. SciBERT is a variant on the BERT (Devlin et al., 2019) model and is pretrained on a scientific corpus of 1.14M articles where 82 percent of the literature comprised of the biomedical domain and the rest was from various computer science domains. In order to extract chemical, protein and disease entities, SciBERT is fine-tuned on different task specific datasets one-by-one, namely, JNLPBA (Collier and Kim, 2004), CHEMDNER (Krallinger et al., 2015) and NCBI Disease Corpus (Dogan et al., 2014) to obtains proteins, chemical and disease annotations respectively.

We use the SciBERT-scivocab-uncased model for NER extraction. The input to the SciBERT model is the pre-processed dataset modified according to the tokenization of BERT. The output of the model consists of the input sentence along with labels according to the BIO scheme where "B" stands for Beginning of an entity tag, "I" stands for Inside of an entity tag and "O" means Outside the entity as can be seen in NER module of Figure 1.

Due to a lack of human gold standard dataset for NER on the CORD-19 data, we do not retain the obtained fine-grained entity annotations. Following the NER tagging, we therefore, tag the Protein, DNA and RNA entities extracted upon fine-tuning the JNLPBA dataset simply as PROTEIN, CHEMDNER as CHEMICAL and NCBI-Disease Corpus as DISEASE. We drop all entities with tags Cell line and Cell type as they could not be merged into any existing categories.

[4] https://spacy.io/api/tokenizer


CORD-19
Sentence level tokenization, word level tokenization , POS tagging
Pre-processed CORD-19
NER Module
Task specific dataset:
• JNLPBA (proteins)
• NCBI (disease)
• CHEMDNER (chemicals)
"...therapy includes Pre-inflammatory cytokines..."
O O B-protein I-protein
CORD-19 with proteins, chemicals and disease mentions
Fine-tuned SciBERT
RE Module
Task specific dataset:
• BC5CDR (chemical-disease relationship)
• CHEMPROT (chemical-protein relationship)
"...therapy includes <e1>Pre-inflammatory cytokines</e1>... " (Label : True)
Entity relations of CORD-19
Fine-tuned SciBERT
KG
Graph Adjacency Matrix
Latent Representation Learning
Link predication using GCN Auto Encoder
Intrinsic Evaluation : Expert score
1. Entity 23
2. Entity 12
3. Entity 80
Ranked list of top similar Entitites with respect to Covid19
Choose the best representation

Pair	Score
(COVID-19, Entity 1)	5
(COVID-19, Entity 2)	3
.	

Figure 1: ERLKG Pipeline

4.3 Relation Extraction

From the NER module we obtain an annotated dataset. To further exploit the underlying information present in the running sentences we perform intra sentence Relationship Extraction (RE) which is the task of identifying relationships between any two named entities present within a sentence. Using this RE module we try to identify the relationships that different pairs of entities have at sentence level. The output from the NER module was further processed in order to discover sentences containing more than one entity. For a given set of entities, E, in a sentence, it is split into $\binom{E}{2}$ instances. So, a single sentence, is represented as: $X = \{e_1, e_2, w_1 ... w_n\}$ where e_1 and e_2 are two tagged entities and w_j is the j^{th} word in the sentence.

An approach similar to the NER module is performed, employing SciBERT for identifying relations from sentences through contextual evidence. We fine tune SciBERT on two datasets, CHEMPROT (Kringelum et al., 2016) and BC5CDR (Li et al., 2016), to capture relations between chemical-protein and chemical-disease pairs.

Following the RE task on the CORD-19 data, we combine the 13 different types of associations obtained upon fine-tuning CHEMPROT as a single relation type called CHEMICAL-PROTEIN. Similarly, only the positive associations obtained upon fine-tuning BC5CDR were retained

Task	Types	# of instances	Total instances
NER Tagged Entities	CHEMICAL	6153	64593
	PROTEIN	42108	
	DISEASE	16332	
Relation Pairs	CHEMICAL -PROTEIN	110485	111916
	CHEMICAL -INDUCED -DISEASE	1431	

Table 1: Statistics of the processed CORD-19 dataset from NER and RE Modules

as CHEMICAL-INDUCED-DISEASE. This ensures that less error is propagated in the absence of gold labels for RE. It also makes sure that the subsequent task of obtaining KG and learning latent entity representations are not misguided during their training phase.

4.4 Knowledge Graph Construction

Statistics of the consolidated set of entity mentions and relation pairs obtained as a result of NER and RE on the CORD-19 dataset can be seen from Table 1. To obtain an overview of the different entities and their association with each other, we generate a KG which is a good association representation of the entire unstructured CORD-19 dataset.

We construct a KG which is defined as $KG = (E, R, G)$, where,

- E: a set of nodes representing disease/ protein/ drug entities
- R: a set of labels representing chemical-protein relation or chemical-disease
- $G \subseteq E \times R \times E$: a set of edges that represent facts connecting entity pairs.

Each fact is a triple $\langle h, r, t \rangle$, where h is the head, r is the relation, and t is the tail of the fact.

4.5 COV19_729

After generating the KG, a list of all entities are supplied to a physician, who clubbed the terms into 3 groups based on their relatedness to COVID-19, i.e., NOT RELATED, PARTIALLY RELATED and HIGHLY RELATED. It was identified that the number of entities in the HIGHLY RELATED group are much less in comparison to the other two categories. Thus, in order to reduce bias, the physician sampled nearly equal number of entities from each group, resulting in a final dataset comprising of 729 entities named as COV19_729. This dataset was then shuffled and passed on to two independent physicians, who provided ratings to each sample indicating how related an entity is to COVID-19 on a scale of 0 (NOT RELATED) to 5 (HIGHLY RELATED).

The inter-rater agreement score (kappa score) is found to be 0.5116, which lies in the moderate agreement range. We, therefore, average out the ratings and propose a relatively large, intrinsic evaluation dataset called COV19_729 for benchmarking COVID-19 related embedding techniques. Table 4 shows a snapshot of the COV19_729.

5 Experiment and Results

5.1 Implementation Details

To generate the intrinsic evaluation dataset, the total list of 78K entities present in our KG are reduced to 5K by removing all entities having less than 5 in-degrees. This is done in order to reduce noise. After experimenting with multiple values, the threshold 5 provided the highest signal to noise ratio. For fine-tuning SciBERT, all hyper-parameters are left at their default values except truncate_long_sequences parameter which is set to false. For training KG embedding, in OpenKE (Han et al., 2018), the dimension is set to 400 and the rest of the parameters are kept as default. In the case of GCN-AE (Kipf and Welling, 2016), for the link prediction task, the learning rate is set to 0.01, epochs to 200, hidden units in the first and second layer as 32 and 16 respectively.

5.2 Link Prediction

Latent entity representation learning of the constructed KG is crucial so that one can effectively analyze associations of any given biomedical entity with respect to COVID-19. Rather than randomly choosing a method, we first evaluate popular KG embedding techniques on a downstream Natural Language Processing task of Link Prediction. We consider Node2Vec (Grover and Leskovec, 2016), Tensor decomposition models like DisMult (Yang et al., 2015) and Geometric models, namely, TransE (Bordes et al., 2013), TransD (Ji et al., 2015), ComplEx (Trouillon et al., 2016) and RotatE (Sun et al., 2019).

The test and validation set is created from the removed edges with the addition of equal number of randomly sampled pairs of false links (nodes that did not have connections in the graph). The test and validation sets have 10 percent and 5 percent of true links, respectively. We use OpenKE (Han et al., 2018) which is an Open-source Framework for Knowledge Embedding techniques. The results are reported based on the model's performance on the test set. The embeddings resulting from these methods are treated as features, along with the graph adjacency matrix and is fed to GCN-AE (Kipf and Welling, 2016). The Average Precision and ROC score of each setting is noted and used to benchmark these embedding types as can be seen in Table 2.

Method	**ROC**	**AP**
RotatE (Sun et al., 2019)	0.858	**0.887**
TransD (Ji et al., 2015)	**0.860**	0.883
TransE (Bordes et al., 2013)	0.853	0.877
DistMult (Yang et al., 2015)	0.855	0.883
ComplEx (Trouillon et al., 2016)	0.852	0.881
Node2Vec (Grover and Leskovec, 2016)	0.821	0.849

Table 2: Link Prediction performance of different KG embedding techniques on test set using GCN-AE

From Table 2 in terms of Average Precision, RotatE performs the best among all KG embeddings while in terms of ROC score, TransD outperforms the rest. Models like TransE capture inversion and composition patterns well, whereas models like DisMult capture symmetrical relationships. But in case of RotatE all the different aspects like symmetry, anti-symmetry, inversion and composition are captured. Also, TransD has a similar performance to RotatE. This is because in our setting every relationship pair has the head and tail entity to be of different entity types (either chemical-protein or chemical-disease). The inherent property of TransD to separate the head and tail entity spaces was useful to model this graph structure. Hence, giving comparable results to RotatE.

Node2Vec performs relatively poor since it relies on the internal mechanism of grouping nodes with identical connection patterns which could be less frequent in our KG as it is not raised from an interaction network and is rather constructed from entities and relations obtained from free text.

5.3 Intrinsic Evaluation

We conduct Intrinsic Evalaution where Table 3 shows the performance of TransD and RotatE embedding methods in terms of Pearson and Spearman correlation scores between the ratings and the cosine similarity scores of entities on the COV19_729 dataset. The cosine similarity scores for each entity was generated with respect to the COVID-19 embedding vector obtained from our proposed pipeline. However, most of the top entities generated by two of our best methods, TransD and RotatE (selected on basis of the link prediction task) were not present in COV19_729 since the said dataset was randomly sampled. In our view, these entities require immediate attention and hence, we conduct another round of scoring to evaluate them and in the process, propose COV19_25.

Entity List	**Spearman Correlation**	**Pearson Correlation**
COV19_729 (TransD)	**0.2186**	**0.2117**
COV19_729 (RotatE)	0.1933	0.1879
COV19_25 (TransD)	**0.4570**	**0.4348**
COV19_25 (RotatE)	0.4240	0.4105

Table 3: Pearson and Spearman Correlation values between the ratings and the cosine similarity scores of 729 randomly sampled entities and 25 pipeline predicted entities with respect to the COVID-19 vector

5.3.1 COV19_25

The top 100 predicted entities from TransD and RotatE were selected and an intersection of the generated entities was taken, which was then passed on to a physician. The physician recommended a list of 25 relevant entities, out of the provided set. This list was then sent to another physician who rated the entities based on their relatedness to COVID-19. This was named as COV19_25.

It is evident from Table 3 that TransD has the highest Pearson and Spearman score on the COV19_729 and COV19_25 datasets. Hence, we use TransD as the final embedding generation method for ERLKG.

6 Discussion

We exploit the contextual evidence from CORD-19 corpus in finding entities and relations. This is followed by KG construction for determining the relatedness between any biomedical entity with respect to COVID-19. A simple co-occurrence ma-

Entity	Tags	Cosine_with_COVID-19	Rating_by_physician
retinoic acid inducible gene-1	protein	-0.079917936	0
hydroxyprolinol	chemical	-0.018277158	2
acute asthma attacks	disease	0.05297136	1
pc18	chemical	0.153728574	1
s1 domain	protein	0.166142751	2
immunodominant epitopes	protein	0.189406748	2
nsp1	protein	0.202800478	3
hcov whole genomes	protein	0.306899184	1
spike glycoprotein	protein	0.413827424	4
receptor binding domain	protein	0.464432383	5

Table 4: Scores given by Raters on few samples from COV19_729. The cosine similarity scores are generated from TransD embeddings

trix based method is not sufficient to capture the different relationship association types. We, therefore, use state-of-the-art SciBERT for the purpose of entity and relationship extraction. We construct a KG from entity pairs and the relationship among them. Our aim was to utilize this KG for effective association analysis for which identifying the best entity representation was necessary. We therefore conduct link prediction task and evaluate popular KG embedding techniques. Since, our KG consists of a simple bare-bone structure, deep learning based KG embedding methods like ConvE were not explored in this work. This is because such methods lead to an increase in the number of hyperparameters while providing little to no explainability.

We face the challenge of an absence of ground truth data for CORD-19 corpus. Thus, we conduct extensive qualitative evaluations and in the process, introduce two gold-standard, annotated entity lists, COV19_25, and COV19_729. COV19_25 consists of 25 entities predicted by the top two embedding techniques, TransD and RotatE while COV19_729 consists of 729 entities sampled from the processed CORD-19 dataset. The ratings were based on an entity's relatedness to COVID-19. From the correlation scores 3 of our intrinsic evaluation, we observe that our model could provide considerable insight in predicting important associations with respect to COVID-19.

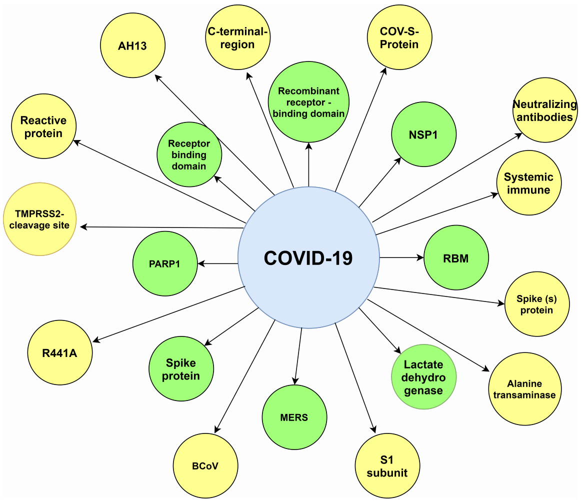


Figure 2: COVID-19 related Proteins based on cosine similarity obtained from ERLKG. Entities color coded Green signify a higher cosine similarity value compared to entities colour coded Yellow.

TransD has the highest Pearson and Spearman score on the COV19_729 and COV19_25 datasets. Hence, we use TransD as the final embedding generation method for ERLKG. Figures 2, 3 and 4 show the top related proteins, chemicals and diseases that ERLKG, using TransD embedding, predicted with respect to COVID-19. Without using any external knowledge resources, our pipeline predicts various chemicals, proteins and diseases that are highly related with COVID-19. These predicted entities could help the biomedical community to get a better understanding of COVID-19. A few top chemicals like Mitoxantrone (Giovannoni et al., 2020), Carfilzomib (Iyer et al., 2020), Flutamide (Cava et al., 2020), Bortezomib (Al Saleh et al., 2020), Lopinavir and Ritonavir (Cao et al., 2020) are being considered as a potential cure for the virus. From the predicted proteins list, entities like PARP1 (Kouhpayeh et al., 2020), Spike protein (Bosch et al., 2003), Lactate Dehydrogenase (Han et al., 2020) and NSP1 (Thoms et al., 2020) have direct relevance with COVID-19. Entities like Ventricular tachycardia (VT) (Wu et al., 2020), Myas-

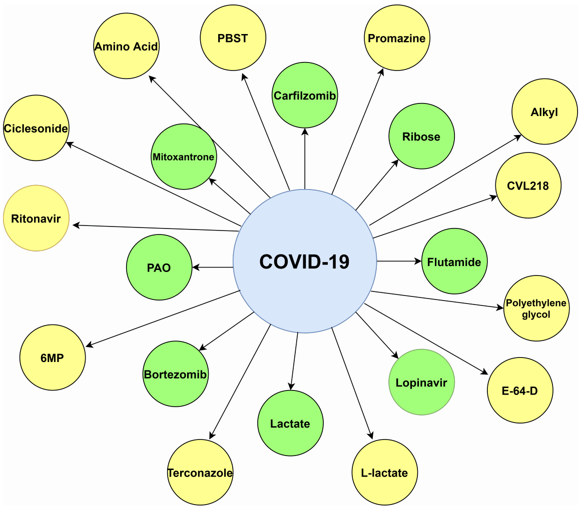


Figure 3: COVID-19 related Chemicals based on cosine similarity obtained from ERLKG. Entities color coded Green signify a higher cosine similarity value compared to entities color coded Yellow.

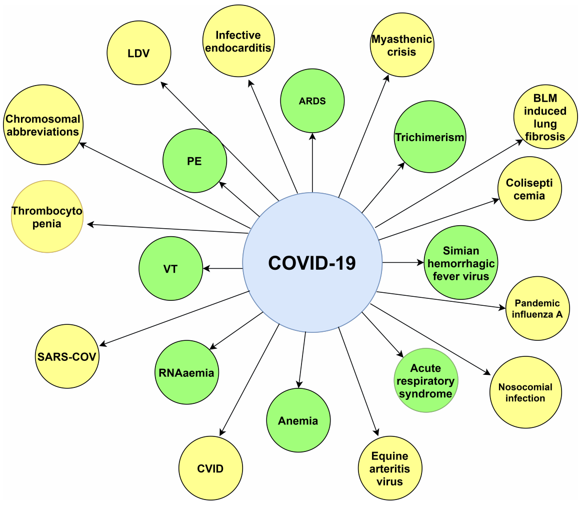


Figure 4: COVID-19 related Diseases based on cosine similarity obtained from ERLKG. Entities color coded Green signify a higher cosine similarity value compared to entities color coded Yellow.

thenic (Delly et al., 2020) crisis, Acute Respiratory Syndrome (Lai et al., 2020), ARDS (Respiratory distress syndrome) (Marini and Gattinoni, 2020) and Thrombocytopenia (Lippi et al., 2020) are a few diseases that are very likely to occur in patients suffering from COVID-19.

7 Conclusion and Future Work

We propose ERLKG, a generic pipeline, for association analysis with respect to a given entity from an unstructured dataset. The part of the pipeline integrating IE and KG construction keeps human-out-of-the-loop. In order to learn the latent representation of the formed KG, we first benchmark various types of KG embedding techniques on the task of Link Prediction. According to our experiments we find TransD and RotatE producing a comparable performance.

In this work our approach is evaluated only on CORD-19 dataset and no additional resources have been employed. However, due to the lack of gold standard data we introduce COV19_729, which is a list of extracted named entities from our pipeline selected randomly and given to physicians for assigning association scores with respect to COVID-19. Owing to random selection most of the entities listed with greater association scores by TransD and RotatE were found to be missing in COV19_729 hence another set was given to physicians from the top entites which we call COV19_25. Finally TransD is used as our best KG embedding technique to predict top entities that are closely associated to COVID-19 from CORD-19 corpus. As a future scope, we plan to implement a normalization and abbreviation expansion module after the detection of entities. The study of these top predicted entities, by the domain experts, can help them understand the different types of associations and relationships they exhibit with respect to COVID-19.

Acknowledgments

The authors acknowledge the Department of Biotechnology, Govt. of India for the financial support for the project BT/COE/34/SP28408/2018. The authors would also like to thank Dr. Shahid Aslam, Department of General Medicine, AMRI Dhakuria Hospitals Kolkata and Dr. Khalid Iqbal, Assistant Professor, Eras Lucknow Medical College and Hospital, for helping with the Intrinsic Evaluation datasets and Sanket Wakade, Department of Design, IIT Guwahati for helping with illustrations. Besides, the authors would like to thank the anonymous reviewers for their valuable comments and feedback.

References

Abdullah S Al Saleh, Taimur Sher, and Morie A Gertz. 2020. Multiple myeloma in the time of covid-19. *Acta haematologica*, pages 1–7.

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical bert embeddings. *ArXiv*, abs/1904.03323.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *EMNLP/IJCNLP*.

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

Berend Jan Bosch, Ruurd van der Zee, Cornelis A M de Haan, and Peter J. M. Rottier. 2003. The coronavirus spike protein is a class i virus fusion protein: Structural and functional characterization of the fusion core complex. *Journal of Virology*, 77:8801 – 8811.

Bin Cao, Yeming Wang, Danning Wen, Wen Liu, Jingli Wang, Guohui Fan, Lianguo Ruan, Bin Song, Yanping Cai, Ming Wei, et al. 2020. A trial of lopinavir–ritonavir in adults hospitalized with severe covid-19. *New England Journal of Medicine*.

Claudia Cava, Gloria Bertoli, and Isabella Castiglioni. 2020. In silico discovery of candidate drugs against covid-19. *Viruses*, 12.

Bin Chen, Xiao Dong, Dazhi Jiao, Huijun Wang, Qian Zhu, Ying Ding, and David J. Wild. 2009. Chem2bio2rdf: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics*, 11:255 – 255.

Chongyan Chen, Islam Akef Ebeid, Yi Bu, and Ying Ding. 2020. Coronavirus knowledge graph: A case study. *ArXiv*, abs/2007.10287.

Nigel Collier and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at jnlpba. In *NLPBA/BioNLP*.

Debsmita Das, Shashank Dubey, Aakash Deep Singh, Kushagra Agarwal, Sourojit Bhaduri, Rajesh Kumar Ranjan, Yatin Katyal, and Janu Verma. 2020. Information retrieval and extraction on covid-19 clinical articles using graph community detection and biobert embeddings.

Fadi Delly, Maryam Jamil Syed, Robert P. Lisak, and Deepti Zutshi. 2020. Myasthenic crisis in covid-19. *Journal of the Neurological Sciences*, 414:116888 – 116888.

Franck Dernoncourt and J. Lee. 2017. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. In *IJCNLP*.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. *ArXiv*, abs/1707.01476.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Daniel Domingo-Fernández, Shounak Baksi, Bruce Schultz, Yojana Gadiya, Reagon Karki, Tamara Raschka, Christian Ebeling, Martin Hofmann-Apitius, and Alpha Tom Kodamullil. 2020. Covid-19 knowledge graph: a computable, multi-modal, cause-and-effect knowledge model of covid-19 pathophysiology. *bioRxiv*.

Gavin Giovannoni, Chris Hawkes, Jeannette Lechner-Scott, Michael Levy, Emmanuelle Waubant, and Julian Gold. 2020. The covid-19 pandemic and the use of ms disease-modifying therapies. *Multiple Sclerosis and Related Disorders*, 39:102073.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Xu Han, Shulin Cao, Lv Xin, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. 2018. Openke: An open toolkit for knowledge embedding. In *Proceedings of EMNLP*.

Yi Han, Haidong Zhang, Sucheng Mu, Wei Wei, Chaoyuan Jin, Yuan Xue, Chaoyang Tong, Yunfei Zha, Zhenju Song, and Guorong Gu. 2020. Lactate dehydrogenase, a risk factor of severe covid-19 patients. *medRxiv*.

Mahalaxmi Iyer, Kaavya Jayaramayya, Mohana Devi Subramaniam, Soo Bin Lee, Ahmed Abdal Dayem, Ssang-Goo Cho, and Balachandar Vellingiri. 2020. Covid-19: an update on diagnostic and therapeutic approaches. *BMB reports*, 53(4):191.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *ACL*.

Thomas Kipf and Max Welling. 2016. Variational graph auto-encoders. *ArXiv*, abs/1611.07308.

Shirin Kouhpayeh, Laleh Shariati, Maryam Boshtam, Ilnaz Rahimmanesh, Mina Mirian, Mehrdad Zeinalian, Azhar Salari-jazi, Negar Khanahmad, Mohammad Sadegh Damavandi, Parisa Sadeghi, et al. 2020. The molecular story of covid-19; nad+ depletion addresses all questions in this infection.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Dong-Hong Ji, Daniel M. Lowe, Roger A. Sayle, Riza Theresa Batista-Navarro, Rafal Rak, Torsten Huber, Tim Rocktäschel, Sérgio Matos, David Campos, Buzhou Tang, Hua Xu, Tsendsuren Munkhdalai, Keun Ho Ryu, S. V. Ramanan, P. Senthil Nathan, Slavko Zitnik, Marko Bajec, Lutz Weber, Matthias Irmer, Saber A. Akhondi, Jan A. Kors, Shuo Xu, Xin An, Utpal Kumar Sikdar, Asif Ekbal, Masaharu Yoshioka, Thaer M. Dieb, Miji Choi, Karin M. Verspoor, Madian Khabsa, C. Lee Giles, Hongfang Liu, K. E. Ravikumar, Andre Lamurias, Francisco M. Couto, Hong-Jie Dai, Richard Tzong-Han Tsai, Caglar Ata, Tolga Can, Anabel Usie, Rui Alves, Isabel Segura-Bedmar, Paloma Martínez, Julen Oyarzábal, and Alfonso Valencia. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7:S2 – S2.

Jens Kringelum, Sonny Kim Kjærulff, Søren Brunak, Ole Lund, Tudor I. Oprea, and Olivier Taboureau. 2016. Chemprot-3.0: a global chemical biology diseases mapping. *Database: The Journal of Biological Databases and Curation*, 2016.

Chih-Cheng Lai, Tzu-Ping Shih, Wen-Chien Ko, Hung-Jen Tang, and Po-Ren Hsueh. 2020. Severe acute respiratory syndrome coronavirus 2 (sars-cov-2) and corona virus disease-2019 (covid-19): the epidemic and the challenges. *International journal of antimicrobial agents*, page 105924.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, D. Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*, 2016.

Giuseppe Lippi, Mario Plebani, and Brandon Michael Henry. 2020. Thrombocytopenia is associated with severe coronavirus disease 2019 (covid-19) infections: a meta-analysis. *Clinica Chimica Acta*.

Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

John J Marini and Luciano Gattinoni. 2020. Management of covid-19 respiratory distress. *Jama*.

Dai Quoc Nguyen, Thanh Vu, T. Nguyen, Dat Quoc Nguyen, and Dinh Q. Phung. 2019. A capsule network-based embedding model for knowledge graph completion and search personalization. *ArXiv*, abs/1808.04122.

Jong Won Park. 2020. Continual bert: Continual learning for adaptive extractive summarization of covid-19 literature. *ArXiv*, abs/2007.03405.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *ArXiv*, abs/1802.05365.

Sumanta Ray, Snehalika Lall, Anirban Mukhopadhyay, Sanghamitra Bandyopadhyay, and Alexander Schonhuth. 2020. Predicting potential drug targets and repurposable drugs for covid-19 via a deep generative model for graphs.

Peter Richardson, Ivan Griffin, C. Tucker, D. Smith, Olly Oechsle, Anne Phelan, Michael Rawling, Edward Savory, and J. Stebbing. 2020. Baricitinib as potential treatment for 2019-ncov acute respiratory disease. *Lancet (London, England)*, 395:e30 – e31.

Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen M. Voorhees, Lucy Lu Wang, and William R. Hersh. 2020. Trec-covid: Rationale and structure of an information retrieval shared task for covid-19. *Journal of the American Medical Informatics Association : JAMIA*.

Andrea Rossi, Donatella Firmani, Antonio Matinata, Paolo Merialdo, and Denilson Barbosa. 2020. Knowledge graph embedding for link prediction: A comparative analysis. *ArXiv*, abs/2002.00819.

Stuart J. Russell and Peter Norvig. 2010. Artificial intelligence : a modern approach - 3rd ed.global ed.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *ArXiv*, abs/1902.10197.

Matthias Thoms, Robert Buschauer, Michael Ameismeier, Lennart Koepke, Timo Denk, Maximilian Hirschenberger, H. Kratzat, Manuel Hayn, T. Mackens-Kiani, Jingdong Cheng, C. Stürzel, T. Fröhlich, O. Berninghausen, T. Becker, F. Kirchhoff, K. Sparrer, and R. Beckmann. 2020. Structural basis for translational shutdown and immune evasion by the nsp1 protein of sars-cov-2. *bioRxiv*.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. *ArXiv*, abs/1606.06357.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Michael Kinney, Ziyang Liu, William. Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020a. Cord-19: The covid-19 open research dataset. *ArXiv*.

Qingyun Wang, Manling Li, X. Wang, Nikolaus Nova Parulian, Guangxing Han, Jiawei Ma, Jingxuan Tu, Ying Lin, H. Zhang, Weili Liu, Aabhas Chauhan, Yingjun Guan, Bangzheng Li, Ruisong Li, Xiangchen Song, Huai zhong Ji, Jiawei Han, Shih-Fu Chang, J. Pustejovsky, D. Liem, A. El-Sayed, Martha Palmer, Jasmine Rah, C. Schneider, and

B. Onyshkevych. 2020b. Covid-19 literature knowledge graph construction and drug repurposing report generation. *ArXiv*, abs/2007.00576.

Xuan Wang, Xiangchen Song, Yingjun Guan, Bangzheng Li, and Jiawei Han. 2020c. Comprehensive named entity recognition on cord-19 with distant or weak supervision. *ArXiv*, abs/2003.12218.

Cheng-I Wu, Pieter G Postema, Elena Arbelo, Elijah R Behr, Connie R Bezzina, Carlo Napolitano, Tomas Robyns, Vincent Probst, Eric Schulze-Bahr, Carol Ann Remme, et al. 2020. Sars-cov-2, covid-19 and inherited arrhythmia syndromes. *Heart Rhythm*.

Bishan Yang, Wen tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. *CoRR*, abs/1412.6575.

Z. Yang, Zihang Dai, Yiming Yang, J. Carbonell, R. Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*.