Original Paper

Large-Scale Biomedical Relation Extraction Across Diverse Relation Types: Model Development and Usability Study on COVID-19

Zeyu Zhang[1,2], PhD; Meng Fang[3], PhD; Rebecca Wu[4], BS; Hui Zong[1,5], PhD; Honglian Huang[1], PhD; Yuantao Tong[1], PhD; Yujia Xie[1], MSc; Shiyang Cheng[1], MSc; Ziyi Wei[1], MSc; M James C Crabbe[6,7,8], PhD; Xiaoyan Zhang[1], PhD; Ying Wang[1,2,3], PhD

[1]Research Center for Translational Medicine, Shanghai East Hospital, School of Life Sciences and Technology, Tongji University, Shanghai, China

[2]Department of Clinical Laboratory Medicine Center, Yueyang Hospital of Integrated Traditional Chinese and Western Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai, China

[3]Department of Laboratory Medicine, Shanghai Eastern Hepatobiliary Surgery Hospital, Shanghai, China

[4]University of California, Berkeley, Berkeley, CA, United States

[5]Institutes for Systems Genetics, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, Sichuan University, Chengdu, China

[6]Wolfson College, Oxford University, Oxford, United Kingdom

[7]Institute of Biomedical and Environmental Science & Technology, University of Bedfordshire, Luton, United Kingdom

[8]School of Life Sciences, Shanxi University, Taiyuan, China

Corresponding Author:
Ying Wang, PhD
Research Center for Translational Medicine
Shanghai East Hospital, School of Life Sciences and Technology
Tongji University
1239 Siping Road
Shanghai, 200092
China
Phone: 86 21 65980233
Fax: 86 21 65981041
Email: nadger_wang@139.com

Abstract

Background: Biomedical relation extraction (RE) is of great importance for researchers to conduct systematic biomedical studies. It not only helps knowledge mining, such as knowledge graphs and novel knowledge discovery, but also promotes translational applications, such as clinical diagnosis, decision-making, and precision medicine. However, the relations between biomedical entities are complex and diverse, and comprehensive biomedical RE is not yet well established.

Objective: We aimed to investigate and improve large-scale RE with diverse relation types and conduct usability studies with application scenarios to optimize biomedical text mining.

Methods: Data sets containing 125 relation types with different entity semantic levels were constructed to evaluate the impact of entity semantic information on RE, and performance analysis was conducted on different model architectures and domain models. This study also proposed a continued pretraining strategy and integrated models with scripts into a tool. Furthermore, this study applied RE to the COVID-19 corpus with article topics and application scenarios of clinical interest to assess and demonstrate its biological interpretability and usability.

Results: The performance analysis revealed that RE achieves the best performance when the detailed semantic type is provided. For a single model, PubMedBERT with continued pretraining performed the best, with an F1-score of 0.8998. Usability studies on COVID-19 demonstrated the interpretability and usability of RE, and a relation graph database was constructed, which was used to reveal existing and novel drug paths with edge explanations. The models (including pretrained and fine-tuned models), integrated tool (Docker), and generated data (including the COVID-19 relation graph database and drug paths) have been made publicly available to the biomedical text mining community and clinical researchers.

Conclusions: This study provided a comprehensive analysis of RE with diverse relation types. Optimized RE models and tools for diverse relation types were developed, which can be widely used in biomedical text mining. Our usability studies provided a proof-of-concept demonstration of how large-scale RE can be leveraged to facilitate novel research.

(J Med Internet Res 2023;25:e48115) doi: 10.2196/48115

KEYWORDS

biomedical text mining; biomedical relation extraction; pretrained language model; task-adaptive pretraining; knowledge graph; knowledge discovery; clinical drug path; COVID-19

Introduction

Background

With the rapid research and development of new biotechnologies, a vast amount of biomedical literature has been published, and biomedical text mining has proved to be invaluable for analyzing the literature, especially for some hot topics that have received much attention from scientific research, such as drug development and public health events. For instance, the COVID-19 pandemic generated a large number of research articles (over 368,000 as of August 3, 2023), making it a priority for the biomedical text mining research community to extract structured knowledge from the corpus of information. The infeasibility of manually reviewing large-scale literature leads to knowledge bottlenecks that contribute to the duplication of research efforts and inefficiency in the development of treatment strategies. The development of better knowledge mining technologies will enable researchers to access relevant knowledge more efficiently and accurately.

Biomedical natural language processing (NLP) techniques are used to extract useful information from biomedical text, such as scientific literature and electronic health records, and these techniques have been extensively employed to extract and retrieve structured information from massive article collections [1]. A critical part of biomedical NLP is relation extraction (RE), which associates a given sentence or text containing entity information with a relation type based on characteristics (entities, context, and semantic features) under predefined categories. RE can be used to implement many application scenarios, such as a structured search and knowledge summarization, and is also the key component for building knowledge graphs (KGs), a powerful way to represent and integrate large-scale textual data to generate new insights [2].

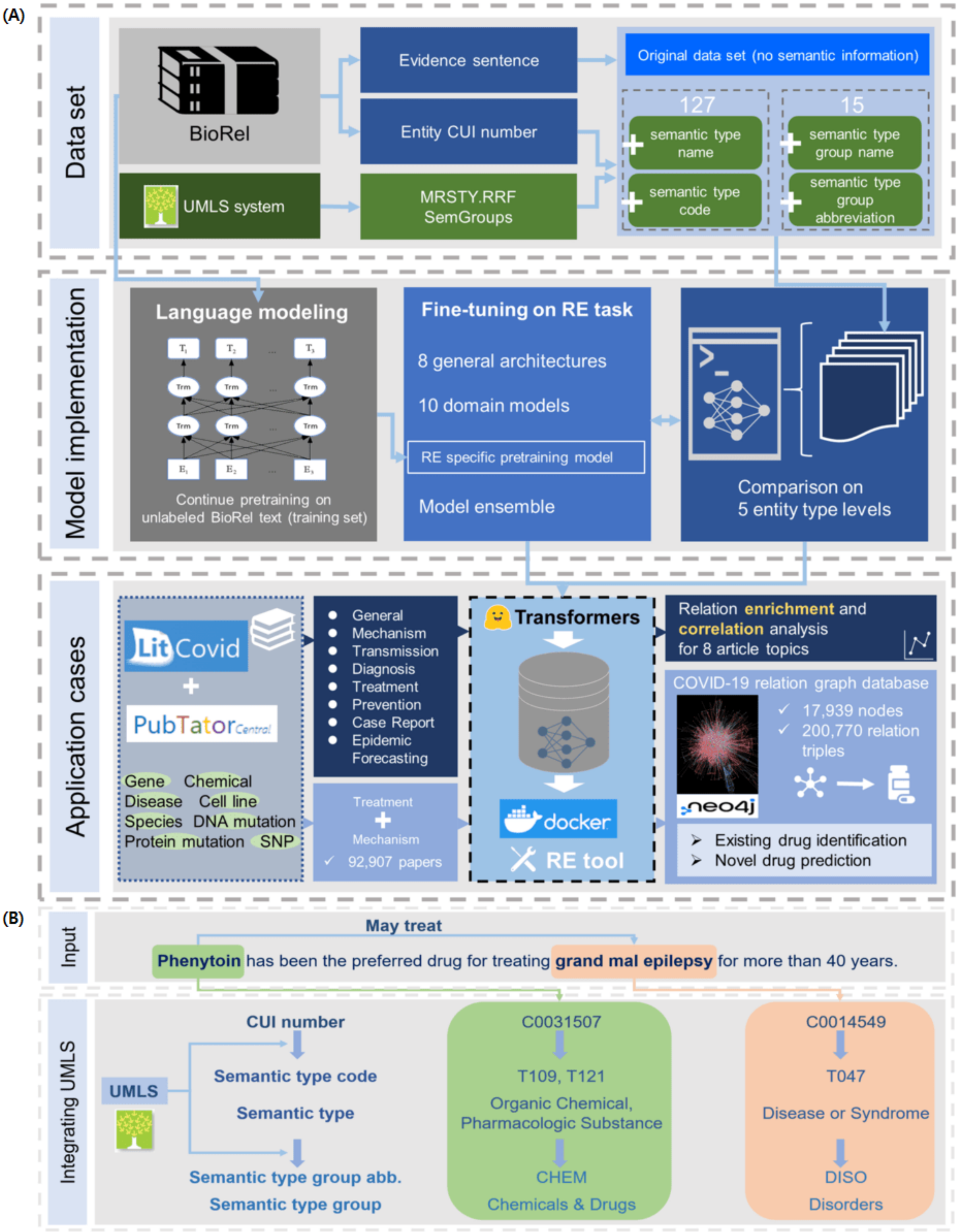
Related Work

Biomedical RE can be performed with a variety of algorithms and data sets. Previous studies adopted computational methods to extract relations, including co-occurrence–based methods, pattern-based methods [3], rule-based methods, feature-based methods [4], and kernel-based methods [5]. For instance, SemRep is a notable rule-based RE tool developed by the National Library of Medicine that can extract semantic relations from sentences in biomedical text [6,7]. Deep neural network–based methods have been shown to achieve better results in various NLP tasks and to have good performance in automatic feature learning [8]. Recent advances in text mining focus on pretrained language models, and many studies have shown that pretrained language models have achieved state-of-the-art NLP methodologies for biomedical text mining [9]. Data sets also play a critical role in deep learning–based RE models, as many RE data sets and models have been built. However, these tasks are mainly focused on the general NLP domain, and the data sets are usually taken from public domain articles, such as Wikipedia [10]. Most existing RE data sets in the biomedical domain, such as GDR [11], are limited in the amount of data and diversity of relations due to labor-intensive manual annotation. BioRel is a large-scale RE data set encompassing 125 biomedical relations via knowledge database utilization and distant supervision [12]. It could facilitate the development of deep learning methods in the extraction of biomedical relations. Although BioRel provides a large-scale data set and has an abundance of biomedical relation types, as far as we know, this data set has not been explored at the entity type level, especially with regard to the influence of the abundance of entity types on RE performance. Moreover, to our knowledge, no research has applied the model with continued pretraining to biomedical RE to build tools and apply them to practical biomedical problems.

Objective

In order to investigate and improve large-scale RE with diverse relation types in biomedical text mining, we conducted a comprehensive study, including a modeling study, tool development, and a usability study with COVID-19 literature. As shown in Figure 1A, we first integrated the Unified Medical Language System (UMLS) [13], one of the most widely used knowledge resources in biomedical NLP, with the BioRel data set to provide richer entity information, generating 4 data sets with different semantic levels of entities. Next, we conducted RE model implementations and evaluated the impact of different levels of semantic information on RE performance. We also compared a variety of pretrained model architectures and domain models, including 8 general pretrained model architectures and 10 domain models, and adopted continued pretraining strategies and ensemble modeling to improve RE performance. The fine-tuned models were integrated into an RE tool as a Docker container. In addition, we investigated biological interpretability and performed multiple application cases on the COVID-19 corpus, including relation enrichment/correlation among topics, relation graph database construction with existing drug identification and novel drug path prediction, non-long/long COVID drug retrieval, and coronavirus-specific relation triple prediction. Our results provide novel insights into biomedical RE in large-scale literature text mining for future studies. The RE task-specific pretrained and fine-tuned models are publicly available at Hugging Face Hub [14-18], and the RE tool is publicly available at Docker Hub [19].

Figure 1. Flowchart of the relation extraction (RE) study and data set construction. (A) Overview flow diagram of 3 main modules in this study. (B) Demonstration of integrating Unified Medical Language System (UMLS) entity information into the data set for a sentence with entity annotation. CUI: Concept Unique Identifier.



Methods

Data Collection and Processing

Many biomedical RE data sets focus on some specific entities and relations, and the data sizes are relatively small. To cover a wide range of biomedical relations and facilitate pretrained model strengths on large data sets, we conducted RE model experiments on BioRel [12], a large-scale data set that includes more than 763,000 sentences (534,000 in the training set) and 69,000 entities. It consists of 125 relations (including “not a relation”) that cover most biomedical relation types, including treatment, side effects, and mechanisms. It also contains relation directions like A “may treat” B, and B “may be treated by” A. Examples of relation types and sentences are shown in

Multimedia Appendix 1. The data set involves multiple entity types, so we matched these entities to the UMLS database using a Concept Unique Identifier (CUI) number to obtain entity semantic information, generating 4 data sets with different entity type information levels. Specifically, we first downloaded UMLS metathesaurus data in rich release format (RRF), which includes all semantic type RRF information with a CUI number, and extracted 127 semantic type names and unique identifiers of semantic types (semantic type codes) by CUI number. Then, we downloaded semantic group data from the UMLS semantic network and extracted 15 semantic type group names and unique identifiers of semantic type groups (semantic type group abbreviations) by semantic type code, which involved the other 2 data sets. Figure 1B shows an example of integrating UMLS entity information to generate 4 data sets for a sentence with a CUI number. So far, we have the following 5 data sets with different entity levels, including the original data set: semantic type name, semantic type code, semantic type group name, semantic type group abbreviation, and no entity type.

In the RE application case, the COVID-19 literature corpus was retrieved from the LitCovid database [20] and the CoronaCentral database [21]. LitCovid is a database (daily updated and curated) for tracking COVID-19 scientific literature that provides article topic categories through manual assignment. The categories are “general information,” “mechanism,” “transmission,” “diagnosis,” “treatment,” “prevention,” “case report,” and “epidemic forecasting.” LitCovid also provides a long COVID collection comprising articles that investigated the persistent, long-term, or delayed symptoms of COVID-19, as well as complications from its treatment at least 4 weeks after the onset of symptoms. We downloaded the LitCovid corpus with PubTator [22,23] detection, which provides automatic entity annotation of 6 biomedical concepts (genes, chemicals, diseases, cell lines, species, and variations). Given the LitCovid categories, we mainly focused on treatment and mechanism literature to extract existing and novel drug therapy knowledge by relation graph database construction. CoronaCentral is a resource providing coronavirus papers containing coronavirus-specific entity recognitions, such as “prevention methods,” “risk factors,” “transmission types,” and “vaccine types.” The coronavirus articles with virus-specific mentions were used to obtain a more specific biomedical association and improve the research community’s understanding of the coronavirus.

Ethical Considerations

This work does not involve any ethical or moral issues. The data used in this study are publicly available dataset and literature data. This study did not meet the criteria for human subject research and a review by an institutional review board was not required.

Language Model Implementation

To evaluate the impact of different levels of semantic information on the performance of RE, we compared the performance of the 4 generated data sets and the original data set based on the same language model. Likewise, we conducted performance analysis on different pretrained model architectures and domain models based on the same data set for semantic level (semantic type name). General pretrained model architectures include ALBERT [24], BERT [25,26], DeBERTa [27], ELECTRA [28], ERNIE [29], LayoutLM [30], RoBERTa [31], and XLNet [32]. For the biomedical domain, we adopted 10 popular models with different pretraining corpus and parameter settings: BioM-ALBERT [33], BioM-ELECTRA [33], PubMedELECTRA [34], BioMed-RoBERTa [35], RoBERTa-PubMed [36], BioBERT [9], SciBERT [37], BioClinicalBERT [38], BlueBERT [39], and PubMedBERT [40].

Moreover, inspired by Gururangan et al [35] (improving performance through task-adaptive pretraining on unlabeled data was found to be effective even after undergoing domain-adaptive pretraining), we proposed a RE-specific continued pretraining strategy to specialize a model and achieve better performance. We adopted the best-performing model to continue pretraining on the unlabeled text (raw text) of the BioRel training set. Continued pretraining was based on masked language modeling (MLM) with the sliding window technique to predict masked tokens in sequence to obtain good contextual understanding, and then, we fine-tuned the RE-specific pretrained model on the labeled training set to evaluate whether performance improved. We also applied an ensemble model integrating the top-performing models to improve the RE performance in this task. The SoftMax function was used to determine the probability that each single model would correctly predict the labels of the relations for input text in the ensemble layer after obtaining the logits of each single model. We employed the ensemble model to produce the final results by using the concept of weighted average to accept the matrix and labels of the submodel prediction scores as input.

Based on the idea of R-BERT [41], tokens of the model input with different entity levels and special separate tokens on both sides were transformed into numerical vectors containing semantic features, which were presented as token embeddings of the model input representation. For the continued pretraining configuring, the ratio of tokens to mask for MLM loss was 0.15 and the fraction for stride in the sliding window was 0.8. We ran 100 epochs with a batch size of 150. For the fine-tuning hyperparameter, the learning rate was set at 2×10^{-5} and the batch size was set at 64, and we ran 5 epochs for each model. Due to the fact that most sentences in the data set were less than 128 words in length, the maximum length of input was set at 128. Finally, labels of relations were output using a fully connected layer. The cross-entropy loss was used as the loss function, and AdamW was used as the optimizer. These language models were implemented with PyTorch (version 1.12.0), Hugging Face Transformers (version 4.26.1) [42], and open source pretraining parameters. Regarding hardware, we conducted pretraining on an NVIDIA GeForce RTX 3090 (24 GB) graphics card. In the fine-tuning phase, we carried out experiments on 2 NVIDIA Tesla P100 (12 GB) systems.

Evaluation Metrics

The models were fine-tuned on the training set and evaluated on the testing set. The testing set included a total of 114,515 samples and covered 125 relation types, and the proportion of each relation type was consistent with the training set. To measure the impact of entity level on biomedical RE tasks and evaluate the performance of language models, we used deep learning metrics that are often adopted: precision, recall, and F1-score. The benchmark score for this task, which combined precision and recall, was the weighted average F1-score. The formulae are presented in Multimedia Appendix 2. These metrics were aggregated across 125 relation types in weighted average levels to compare the overall performance. Precision-recall curves were drawn for 5 entity levels and all language models.

Usability Study With the COVID-19 Corpus

To demonstrate the biological significance and interpretability of the RE application, we conducted experiments to examine relation type enrichment and correlation within different article topics. Equal numbers of articles covering 8 topics from LitCovid were filtered out separately (downloaded on March 7, 2023), and then, we detected relations from these 8 sets to compare relation enrichment between topics. The correlation index between topics by diversity and amount of relation was calculated to illustrate the strength of the correlation between every 2 topics. Principal component analysis (PCA) was also performed to display the distance between the 8 topics based on relation type.

To understand the clinical treatment and drug discovery for COVID-19, based on the KG concept, we retrieved “treatment” and “mechanism” research from LitCovid to build a relation graph database. Sentences with entities were used in RE prediction, and the results were encoded in triple format (ie, head node–relation–tail node), producing metadata regarding the nodes and their relational connections, as well as the correlating origins and context. The triple data mainly contained 2 entities and their relation, including the direction of the relation. We converted these triple data and imported them into Neo4j, a popular graph database management system, using custom Cypher scripts.

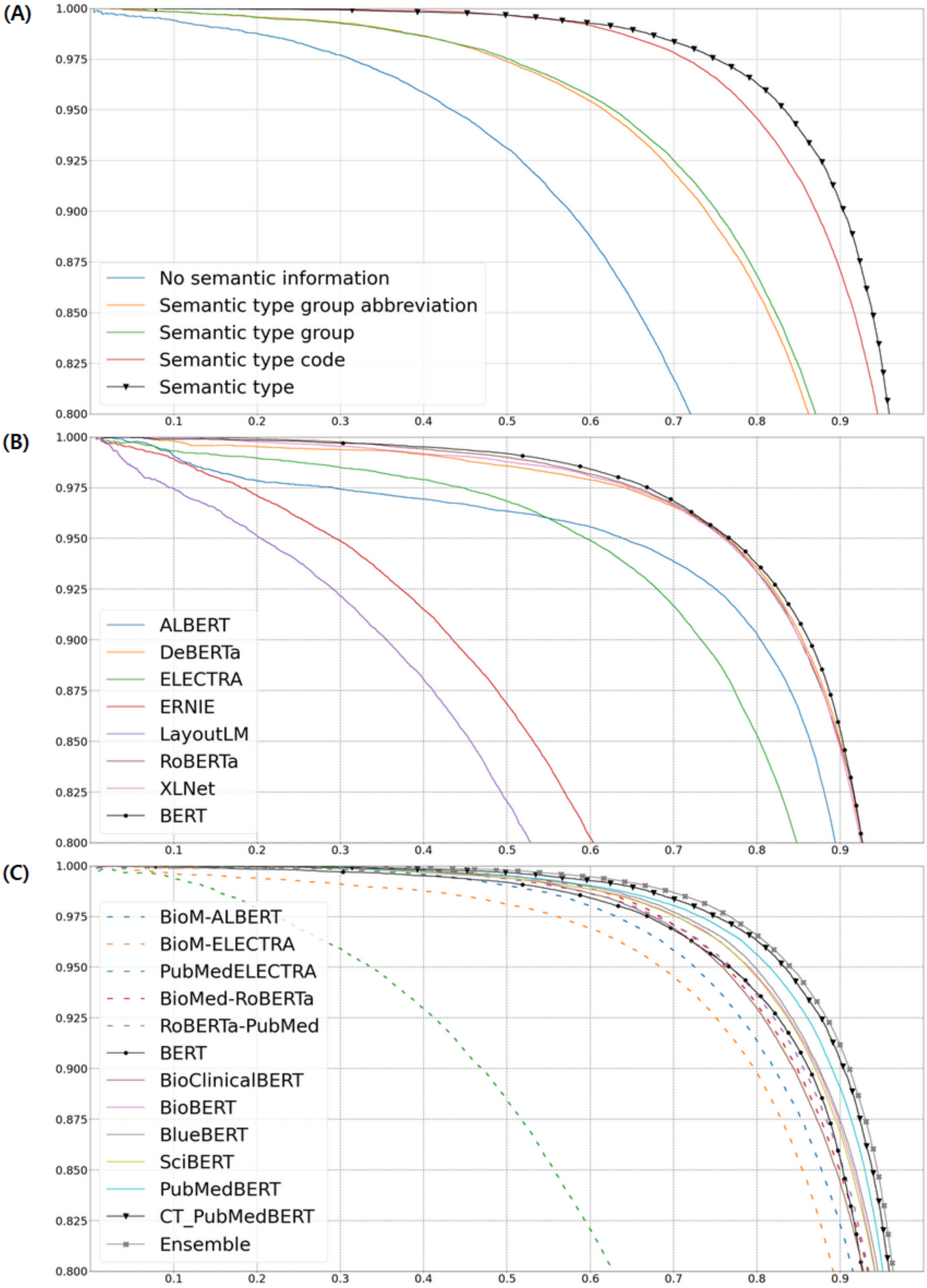
Moreover, we used the relation graph database to identify existing drugs and discover novel drug candidates. For existing drugs, we defined 3 conditions to identify drugs with more plausible associations with COVID-19, which also helped us to visualize the paths associated with the drug-disease. To discover potential drugs not directly connected to COVID-19, we adopted the drug paths defined by Sosa et al [43]. They used the Global Network of Biomedical Relationships (GNBR) [44], a large KG, to predict new drug paths for rare diseases, and they defined three 4-node paths connecting the drug-disease. Within path a, the medication uses the same genetic process to treat a different condition. Within path b, the medication addresses a condition that is treated similarly to the condition of interest. Within path c, the medication addresses the condition through 2 connected genes. We also added additional qualifications on the 3 drug paths based on the original pattern to find more plausible nodes and edges. The results were ranked by the Adamic Adar score [45]. This algorithm was used to compute the closeness of nodes based on their shared neighbors. The calculation of the Adamic Adar score is shown in Multimedia Appendix 2.

Results

Impact of Entity Information Level and Language Model Implementation

Data sets with 5 different entity levels were used for experiments, and they were fine-tuned with the same pretrained model (PubMedBERT with continued pretraining) for comparison. The precision-recall curves are shown in Figure 2A. The performance of RE was the best when using semantic name (F1-score=0.8998) and semantic name code (F1-score=0.8817), the next best when using semantic type group (F1-score=0.8279) and semantic type group abbreviation (F1-score=0.8230), and the worst when using the original data set that had no entity information (F1-score=0.7440). The performance values for semantic name were slightly higher than for semantic name code, and were slightly higher for semantic type group than for semantic type group abbreviation. The results showed that the detailed semantic name can provide more semantic information than the abbreviation and code name to improve the performance of RE. Multimedia Appendix 3 shows the F1-score difference of each relation type for the 4 entity information levels compared to the no semantic type. It can be seen that “semantic type group abbreviation” had 104 relation types with improvement, 13 with no change, and 8 with a decrease; “semantic type group” had 112 relation types with improvement, 10 with no change, and 3 with a decrease; “semantic type code” had 115 relation types with improvement, 7 with no change, and 3 with a decrease; and “semantic type” had 125 relation types with improvement. Regarding the specific performance of each relation type, there were some relation types with substantial performance improvement at all 4 levels, such as “gene encodes gene product” and “process involves gene,” as these relation types tended to concentrate on certain fixed combinations. Similarly, when using “semantic type” and “semantic type code,” some relation types were further enhanced compared to “semantic type group” and “semantic type group abbreviation,” such as “associated with malfunction of gene product” and “has physiologic effect,” indicating that these relation types were more focused on the combination of entity types. In addition to the above-mentioned relation types that were affected by entity type combination, there were also relation types whose performances were affected by the semantic features behind the entity type information, resulting in “semantic type group” and “semantic type” performing better than “semantic type group abbreviation” and “semantic type code,” respectively, such as “disease excludes primary anatomic site” and “is abnormal cell of disease,” indicating that the full texts of “semantic type group” and “semantic type” were transformed into numerical vectors containing semantic features during tokenization, such that the model input representation had richer token embeddings, which was helpful for the model to identify the relation type. Therefore, entity semantic information helps RE from not only the combination of entity types, but also its semantic features in token embeddings.

Figure 2. Precision-recall curves (x-axis: recall, y-axis: precision). (A) The performance comparison of 5 entity levels. (B) The performance comparison of 8 pretrained model architectures. (C) The performance comparison of 10 domain models, CT_PubMedBERT, and the model ensemble (BERT included for comparison). BERT-based domain models are represented by solid lines, and nonBERT-based domain models are represented by dashed lines.



We implemented a variety of pretrained model architectures and domain models. The results are shown in Figure 2B, Figure 2C, and Table 1. We used the precision-recall curve to visualize performance comparison, and the precision, recall, and F1-score to evaluate the performance of each pretrained model. For general model architecture, the performance of BERT was slightly better than that of DeBERTa, RoBERTa, and XLNet, achieving an F1-score of 0.8790. For domain models, PubMedBERT with our continued pretraining (CT_PubMedBERT) achieved the best results for a single language model, with a weighted average precision of 0.8998, recall of 0.9010, and F1-score of 0.8998. The ensemble model integrating BioBERT, BlueBERT, and CT_PubMedBERT with continued pretraining (the 3 best-performing single models) achieved the best overall results with an average precision of 0.9026, recall of 0.9046, and F1-score of 0.9028.

Table 1. Precision, recall, and F1-score of different architectures and models.

Architecture and model	Precision	Recall	F1-score
ALBERT			
ALBERT	0.852	0.8566	0.8533
BioM-ALBERT	0.8611	0.8625	0.8577
BERT			
BERT	0.8784	0.881	0.879
BioClinicalBERT	0.8675	0.8718	0.8684
BioBERT	0.8836	0.8857	0.8838
SciBERT	0.8808	0.8834	0.8815
BlueBERT	0.8833	0.8865	0.8838
PubMedBERT	0.892	0.894	0.8925
CT_PubMedBERT	0.8998	0.901	0.8998
Ensemble[a]	0.9026	0.9046	0.9028
DeBERTa			
DeBERTa	0.8757	0.8775	0.8763
ELECTRA			
ELECTRA	0.8231	0.8249	0.8236
BioM-ELECTRA	0.848	0.8508	0.8487
PubMedELECTRA	0.6904	0.7104	0.6916
ERNIE			
ERNIE	0.6753	0.7007	0.6798
LayoutLM			
LayoutLM	0.6287	0.6645	0.6311
RoBERTa			
RoBERTa	0.8736	0.8752	0.874
BioMed-RoBERTa	0.8685	0.874	0.8696
RoBERTa-PubMed	0.8752	0.878	0.8758
XLNet			
XLNet	0.8749	0.8762	0.8754

[a] The ensemble model integrated BioBERT, BlueBERT, and CT_PubMedBERT with continued pretraining.

The F1-scores for each relation label in CT_PubMedBERT are shown in Multimedia Appendix 4. There were 48 relation labels (87,369/114,565, 76.3% of the total testing set) that had F1-scores greater than 90%, and the categories that performed the best were “biological process involves gene product,” “chemical or drug initiates biological process,” “chemotherapy regimen has component,” “gene encodes gene product,” “gene product encoded by gene,” “gene product has organism source,” “gene product plays role in biological process,” “is component of chemotherapy regimen,” “is organism source of gene product,” and “organism has gene.” These categories all demonstrated a precision, recall, and F1-score of 99%. Multimedia Appendix 5 shows the F1-score change of CT_PubMedBERT compared with PubMedBERT for each relation type. We can see that CT_PubMedBERT improved most relation types (improved 88 relation types, caused no change in 6 relation types, and decreased 31 relation types). Specifically, CT_PubMedBERT showed improvement over PubMedBERT for relation types with low performance owing to lack of data (the proportion of the relation type in the data set reflects its probability of occurrence in real scenarios), and 13 of the 15 relation types with F1-scores of less than 0.5 showed improvement (1 had no change and 1 decreased), indicating that the task-specific language modeling of CT_PubMedBERT has advantages for identifying minority relation types.

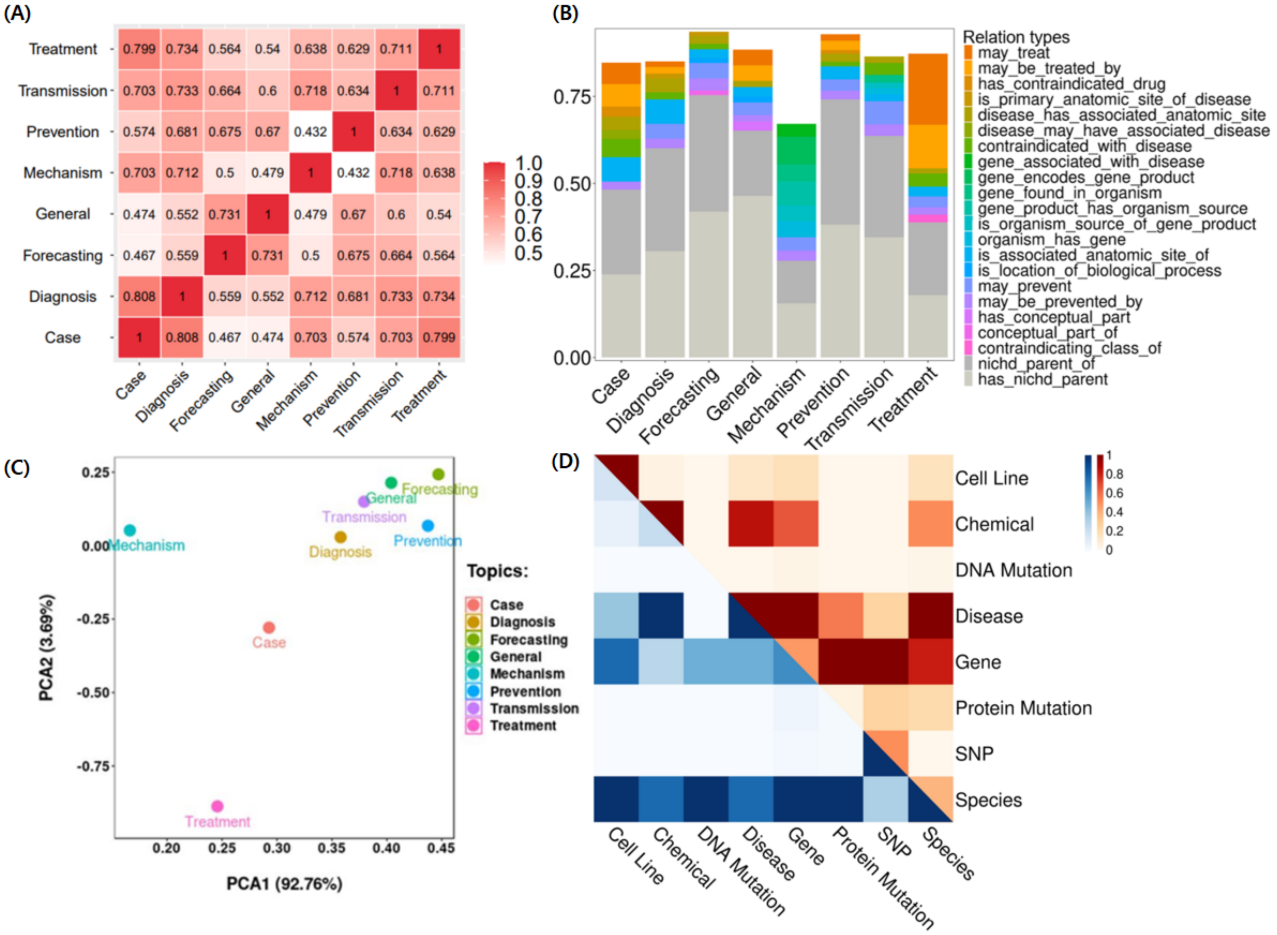
We set the epoch to 100 to improve visualization of the F1-score trend and training step loss for our large data set and complex relation types. As shown in Multimedia Appendix 6, the F1-score increased continuously, and the loss continued to decrease even after 100 epochs. The focal loss in the training set converged more slowly, but still fluctuated and decreased slightly. The model did not fit the large data set perfectly even after 100 epochs. Nonetheless, in terms of running time and performance evaluation, the results were satisfactory after just 5 epochs.

Owing to complicated and heavy deep learning dependencies that are difficult to install and launch, we compiled a Docker image with all the relevant frameworks for this RE model. It contained all the required libraries, packages, and files, saving time for framework installation. The usage tutorials and documentation for the interactive Jupyter Notebook are provided on the Docker hub [19], and the RE tool is publicly available as a Docker container online. We also uploaded our continued pretraining model and its fine-tuned RE models to the Hugging Face Hub [14-18].

Usability Analysis With COVID-19 Cases

The biological significance of RE was evaluated by correlation and relation enrichment. Figure 3A shows the pairwise correlation between 8 topics calculated by relations, demonstrating the strength of the correlation between 2 topics from the perspective of RE. The correlations between "case report" and "diagnosis," and "case report" and "treatment" were relatively high, while the correlations between "mechanism" and "prevention," and "mechanism" and "general" were relatively low. Macroscopically, descriptions related to "diagnosis" and "treatment" of patients were often included in many "case report" articles, while "mechanism" research of viruses often investigated specific biological processes and gene activities, and it did not involve "general" descriptions and "prevention" information. Therefore, the correlation coefficient calculated from RE conformed to the actual content overlap between different topics. We also compared the proportions of the top 10 relation types in each of the 8 article types. Figure 3B shows that in the "treatment" topic, the proportion of treatment-related relation types was relatively high, while in the "mechanism" topic, the proportion of relation types related to gene function was relatively high, which shows that the RE model had biological significance. Specifically, Figure 3C shows the PCA of relation type for 8 topics. The distance between 2 topics represented their relation-based similarity, and articles involving "treatment," "mechanism," and "case report" had relatively topic-specific relation types compared with other topics. We also performed visualization of relation strength between different entity types by the number of relations and relation type, as shown in Figure 3D. "Gene," "disease," and "chemical" had strong RE associations with other entity types, as did "species," owing to the inclusion of "SARS-CoV-2," and these findings are in line with real scenarios.

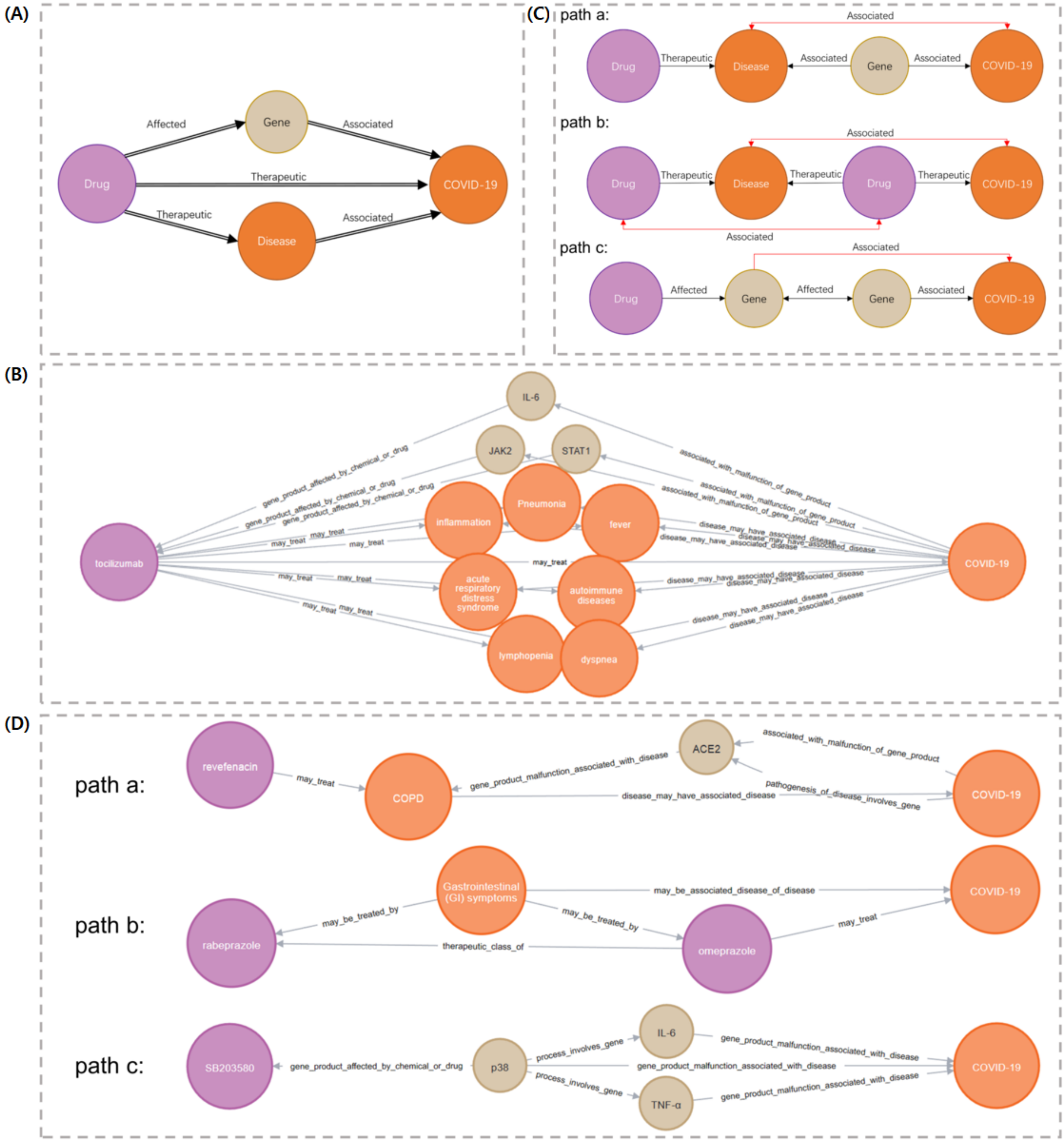
Figure 3. Visualization of topics and entities based on relation extraction (RE). (A) Pairwise correlation between 8 topics calculated by relations. (B) Proportion of the top 10 relations among 8 topics. (C) Principal component analysis (PCA) of 8 topics. (D) Strength of correlations between different types of entities calculated by RE. The top right part is calculated by the number of relations, and the bottom left part is calculated by the number of relation types. SNP: single-nucleotide polymorphism.



We built a COVID-19 relation graph database composed of mechanism and treatment research. The abstracts were downloaded on June 12, 2023, and the RE pipeline was performed. A total of 1,849,915 relation triples were identified in 92,907 papers. After merging identical triples and removing "not a relation" triples, a total of 17,939 unique entities and 200,770 unique relation triples were imported into the Neo4j graph database (version 4.4.21), with node attributes containing entity names and normalized IDs, and edge attributes containing the relation type, the direction, the number of sources, and the mean and sum of the probabilities. The relation graph database constituted a coherent network and allowed entities and relations to be queried, browsed, and navigated. The visualization made it possible for us to view and analyze the network by filtering nodes or edges, and allowed us to explore paths from one piece of information to another. Moreover, it allowed us to add updated information and analyze its indications against the relation graph database. Multimedia Appendix 7 shows a demonstration of the relation graph database. The orange nodes are diseases, the brown nodes are genes, and the purple nodes are chemicals or drugs. Owing to the data sources, the causation relations from "treatment" papers predominantly reflected COVID-19 pathologies and therapies, and those from "mechanism" papers mainly reflected molecular interactions and biological processes. For example, as shown in Multimedia Appendix 7, we can visualize how topotecan acts as a potential drug through the relation graph database (topotecan reduces SARS-CoV-2–induced inflammation by affecting human topoisomerase 1 [46,47]). The importable data of the relation graph database for Neo4j are publicly available [48].

The relation graph database was notable for its extensive coverage of drug-disease interactions, as well as the associated nodes of biological mechanisms linked to COVID-19. As shown in Figure 4A, to identify existing drugs having more plausible associations with COVID-19, 3 conditions must be met. First, the drug should be able to produce an effect on a gene associated with COVID-19. Second, the drug should be therapeutic for a disease associated with COVID-19. Third, the above relations should be identified in at least two different texts to ensure plausibility. The top 3 results, ordered by the Adamic Adar score, were tocilizumab, ritonavir, and baricitinib. The graph for tocilizumab (ranked first) is shown in Figure 4B. Tocilizumab mainly affects interleukin 6 (IL-6), Janus kinase 2 (JAK2), and signal transducer and activator of transcription 1 (STAT1) [49,50], while it has a therapeutic effect on COVID-19–related inflammation, dyspnea, and autoimmune diseases [51-53].

Figure 4. Existing and novel drug graphs. (A) Existing drug identification patterns. (B) Existing drug identification examples. (C) Novel potential drug prediction patterns. (D) Novel potential drug examples. ACE2: angiotensin converting enzyme 2; COPD: chronic obstructive pulmonary disease; IL: interleukin; JAK2: Janus kinase 2; STAT1: signal transducer and activator of transcription 1; TNFα: tumor necrosis factor-α.



To discover novel potential drugs, we used the relation graph database to retrieve 3 drug paths. As shown in Figure 4C, the first node drug in paths a, b, and c is not directly connected to COVID-19 in the entire relation graph database but is connected to COVID-19 through 2 intermediate nodes. The black edges represent the interentity connections defined by Sosa et al [43], and the red edges are the edge qualifications we performed. In all 3 paths, the second node (disease or gene) must be associated with COVID-19 (a disease can represent a symptom associated with COVID-19 or a disease that occurs concurrently with COVID-19). In path b, we also defined that there must be an association between the 2 drugs. The results of the 3 paths were finally ranked by the Adamic Adar score, where the results of path a contained 49 chemicals, the results of path b contained 97 chemicals, and the results of path c contained 9 chemicals. The list of results and the entire graphs of the 3 paths (JSON format) are available online [54]. As shown in Figure 4D, we also provide an example for each of the 3 paths. In path a, revefenacin was a drug for chronic obstructive pulmonary disease (COPD), which can help relax the lung muscles and help relieve cough and shortness of breath [55], while COPD has many potential negative interactions with COVID-19 [56] and abnormal expression of angiotensin converting enzyme 2 (ACE2) plays an important role in both COPD and COVID-19 [57,58]. Since revefenacin and COVID-19 did not appear together in all literature abstracts, we did a full-text search review and found that Djokovic et al [59] used structure-based molecular modeling and physiological-based pharmacokinetic modeling for drug repurposing, and the full text mentions revefenacin as a candidate with potential activity on the SARS-CoV-2 main protease. In path b, we focused the second node of the disease on a specific symptom (gastrointestinal symptom), and rabeprazole and omeprazole have been used to treat gastrointestinal diseases and have the same type of efficacy [60,61], while omeprazole has been used to treat COVID-19 [62]. Rabeprazole was also mentioned in the full text of the study by Ray et al [63] as a possible treatment for COVID-19, either alone or in combination with other drugs. In path c, SB203580 can affect p38 [64], while abnormalities in p38, IL-6, and tumor necrosis factor-α (TNFα) have all been shown to be associated with COVID-19 [65-67]. In addition, p38 has also been shown to be associated with IL-6 and TNFα activity [68]. We also found that the potential treatment of inflammation with SB203580 as a p38 inhibitor has been discussed in detail in the full text of the review on COVID-19 by Malekinejad et al [69].

We also filtered out long COVID articles and non-long COVID articles of "treatment" separately and extracted drugs by detecting therapeutic relations. As shown in Multimedia Appendix 8, the results based on RE indicated that 107 drugs were involved in the treatment of long COVID, 154 drugs were involved in the treatment of non-long COVID, and only 47 drugs appeared for both non-long COVID and long COVID. For the CoronaCentral data (downloaded on March 27, 2023) [21], we retrieved coronavirus-specific entity types, including viral lineages, risk factors, symptoms, and prevention methods, along with general entities, like drugs and diseases, from all SARS-CoV-2 articles to extract relations between them. Finally, we built a set of SARS-CoV-2–specific knowledge triples. Benefiting from these coronavirus-specific entity types, the RE was applied to extract more detailed results. The CoronaCentral data have been made available [70].

Discussion

Principal Findings

In this study, we comprehensively investigated biomedical RE from the perspectives of data, model, and application. The study conducted performance analyses for different entity information levels, pretrained model architectures, and domain models. We also proposed a continued pretraining model for the specific RE task, which achieved the best performance for a single model. The RE models were then integrated as a convenient Docker tool with applications to practical biomedical problems. LitCovid was used to obtain a corpus of literature sorted by topic. Relation enrichment and correlation analysis for 8 types of topics demonstrated that there were differences in relation type between article topics, indicating the biological significance of text mining from the perspective of RE. For articles on treatment and mechanism topics, the output relation triples between key entities were constructed as a relation graph database, which not only allowed us to obtain known therapeutic drugs from existing research, but also helped us to perform drug repurposing via link algorithms and predict novel drug paths. In addition, RE on the treatment corpus of non-long COVID and long COVID could help us to pinpoint the therapeutic drug differences between them. In order to apply RE more profoundly, we also extracted relations between coronavirus-specific entities, like symptoms, viral lineages, and risk factors, from CoronaCentral, giving us a more precise knowledge network of the coronavirus.

The data set we applied consisted of 125 biomedical relations covering treatment, components, side effects, metabolic mechanisms, etc. All relation types and example sentences are presented in Multimedia Appendix 9. Compared to existing biomedical data sets, such as ChemProt [71,72], DDI [73], and GAD [74], data sets integrating BioRel and UMLS contain significantly more words, entities, and relations. The experiments on data sets with different entity levels showed that relation prediction was the most accurate when the input contained the semantic type name of the entity. Additionally, the entity types with a unique identifier (such as "T047" and "DISO," which appear in code and abbreviation formats) performed worse than those with a full expression, and could result in the wrong prediction. Finally, although semantic type group can categorize all semantic types, the information is also more likely to lead to an incorrect prediction. Taken together, these results suggest that more detailed information on entity types with semantic information can help to significantly improve the accuracy of RE.

In model comparison, CT_PubMedBERT achieved the best performance, and biomedical domain models had better performance than general models. The reason was that the model pretrained with biomedical domain-specific literature corpora and unlabeled task-specific data sets had the most extensive background information, giving it a more precise handle on the meaning of individual words. Domain models also integrated the contextual information of sentences into the word vector owing to the use of domain-specific vocabulary and pretraining from scratch (as opposed to the Word2Vec model [75,76]). Moreover, the ensemble model improved the overall RE performance, as confirmed by the performance indicators.

In practical applications, different data sources (literature) may lead to controversial statistical analysis results. For example, hydroxychloroquine has been increasingly found to be useless in the treatment of COVID-19 [77], but many studies, especially early stage studies, include a description that it is a potential and safe drug [78-80]. Therefore, data preprocessing is critical to the quality of RE results. For example, to find more promising drug–COVID-19 treatment triples, 3 preprocessing steps were used. First, papers within the last year (June 1, 2022, to June 1, 2023) were selected to remove studies in the early stage of the pandemic that were not in-depth. Second, the Altmetric score of each paper was crawled, which is an important indicator of research attention, and papers with higher-than-average scores (about 10% of all papers) were used. Third, a rule-based approach was used to increase credibility by selecting text with conclusive descriptions in the abstract. According to these criteria, the top 3 drugs ranked by the sum of the probability scores included ritonavir, dexamethasone, and baricitinib, which are currently widely used and validated in the treatment of COVID-19 [81].

Biomedical and clinical researchers typically keep track of new discoveries through extensive collections of scientific articles, and language model–based NLP techniques are of great help to researchers in extracting information of interest [82]. Research into KG construction, graph path prediction, and automatic text summary greatly benefits from the automated RE process [83,84]. Building a KG can be very tedious and time-consuming if the entities and relations need to be manually identified and inputted. The RE models and tools we built improve the development of large-scale biomedical RE and enable the automatic extraction of relations from scientific articles. These tools make it possible to rapidly build and update a KG. More importantly, applying graph algorithms to KGs enables knowledge discovery, and the representations of KGs can be used for many downstream tasks. The large-scale RE in biomedical text mining can help inform future research, and rigorous conclusions can be drawn through further experiments.

The statistics of the error analysis involved the CT_PubMedBERT model mispredicting the gold standard relation type as other relation types. Multimedia Appendix 10 shows the top 100 misprediction types by percentage of total errors, and the mispredictions can be mainly divided into 3 categories. The first category included mispredicting the direction of the relation, such as the first (7.03%) and second (5.95%) ranked misprediction types (“anatomic structure is physical part of” and “has physical part of anatomic structure”), which both express anatomic structure part relations but in different directions. The second category included confusing relation types involving subordinate meanings with “not a relation,” such as “nichd parent of” and “chemical structure of,” indicating that the model may be less effective in distinguishing subordinate relation types from no relation. The third category included errors in distinguishing similar relation types, such as “disease has primary anatomic site” and “disease has associated anatomic site,” both of which mean that the disease has an anatomic site. Therefore, the improvement of these 3 categories of errors is a direction for future research, for example, data augmentation may improve the second and third categories, while manual review to generate custom rules may improve the first category.

Limitations

Our study provides a biomedical RE implementation that makes analysis accessible to the research community. Nevertheless, several limitations exist. First, the rare entity type and unbalanced data distribution of relation categories limit the performance of the RE. For example, “associated with malfunction of gene product,” “biomarker type includes gene product,” and “gene product has structural domain or motif” had a relatively small amount of data but achieved F1-scores above 0.95 because semantic features in these relations are usually unambiguous and distinctive. The relations “anatomic structure is physical part of,” “disease has associated anatomic site,” and “may treat” achieved comparatively high F1-scores owing to the greater amount of data. In addition, we observed multiple errors and reversed mistakes between relations. Our RE model would benefit the most from optimizations to the data set and algorithm. Short sentences include fewer words and might not provide the RE model with detailed information for correct prediction. Domain-specific expressions, such as formula symbols, measurement units, and proper nouns, are frequently used in scientific writing. Therefore, incorporating customized rules into biomedical RE will increase the efficiency of the model. To improve performance by enhancing semantic features, effective data augmentation might also be used for relation types with small data sizes.

Conclusions

Our study broke new ground in the pretrained language model it used, the comprehensiveness of its biomedical RE topics, the many types of relations it covered, and the insights it generated into hot and prolific scientific topics. We not only evaluated the impact of entity semantic richness, but also compared different model architectures and domain models. We also proposed a continued pretraining model for our specific RE task and fine-tuned it to achieve the best performance. The RE models were packaged as an easy-to-use tool and were applied to the COVID-19 corpus for usability analysis. Furthermore, our relation graph database pipeline can be applied to other large-scale biomedical text mining areas of interest and is not intrinsically limited to cases as shown in this study. It is our hope that our contributions to the field of knowledge mining and the presented tools will facilitate other biomedical and clinical research in the future.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (81972914 and 81573023), the Innovation Group Project of Shanghai Municipal Health Commission (2019CXJQ03), the Fundamental Research Funds for the Central Universities (22120200014), and the Shanghai “Rising Stars of Medical Talent” Youth Development Program (2019-72).

Data Availability

The following data have been made available: Neo4j data of the COVID-19 relation graph database [48], novel drug path data [54], and CoronaCentral data [70].

Authors' Contributions

ZZ was involved in methodology, investigation, and analysis, and drafted the manuscript. MF, RW, HZ, YT, HH, YX, SC, ZW and YW provided technique assistance. RW, YW, and MJCC checked the results and revised the manuscript. YW and XZ obtained funding, supervised the data analysis, and coordinated the overall study. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Examples of relation types and sentences.
[DOCX File , 25 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Formulae used in the study.
[DOCX File , 24 KB-Multimedia Appendix 2]

Multimedia Appendix 3

F1-score difference of each relation type for the 4 entity information levels compared to the no semantic type.
[DOCX File , 36 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Performance of CT_PubMedBERT for each relation type.
[DOCX File , 38 KB-Multimedia Appendix 4]

Multimedia Appendix 5

F1-score difference of CT_PubMedBERT compared with PubMedBERT for each relation type.
[DOCX File , 36 KB-Multimedia Appendix 5]

Multimedia Appendix 6

F1-score and loss trend for 100 epochs.
[DOCX File , 173 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Relation graph database visualization.
[DOCX File , 703 KB-Multimedia Appendix 7]

Multimedia Appendix 8

Venn diagram of drugs between non-long COVID and long COVID.
[DOCX File , 185 KB-Multimedia Appendix 8]

Multimedia Appendix 9

All relation types and example sentences.
[XLSX File (Microsoft Excel File), 25 KB-Multimedia Appendix 9]

Multimedia Appendix 10

Top 100 misprediction types by percentage.
[DOCX File , 30 KB-Multimedia Appendix 10]

References

1. Fiorini N, Leaman R, Lipman DJ, Lu Z. How user intelligence is improving PubMed. Nat Biotechnol 2018 Oct 01:4267 [doi: 10.1038/nbt.4267] [Medline: 30272675]
2. Nicholson DN, Greene CS. Constructing knowledge graphs and their biomedical applications. Comput Struct Biotechnol J 2020;18:1414-1428 [FREE Full text] [doi: 10.1016/j.csbj.2020.05.017] [Medline: 32637040]

3. Corney DPA, Buxton BF, Langdon WB, Jones DT. BioRAT: extracting biological information from full-length papers. Bioinformatics 2004 Dec 22;20(17):3206-3213 [doi: 10.1093/bioinformatics/bth386] [Medline: 15231534]
4. Kim S, Liu H, Yeganova L, Wilbur WJ. Extracting drug-drug interactions from literature using a rich feature-based linear kernel approach. J Biomed Inform 2015 Jul;55:23-30 [FREE Full text] [doi: 10.1016/j.jbi.2015.03.002] [Medline: 25796456]
5. Zhang Y, Lin H, Yang Z, Wang J, Li Y. A single kernel-based approach to extract drug-drug interactions from biomedical literature. PLoS One 2012;7(11):e48901 [FREE Full text] [doi: 10.1371/journal.pone.0048901] [Medline: 23133662]
6. Kilicoglu H, Rosemblat G, Fiszman M, Shin D. Broad-coverage biomedical relation extraction with SemRep. BMC Bioinformatics 2020 May 14;21(1):188 [doi: 10.1186/s12859-020-3517-7] [Medline: 32410573]
7. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform 2003 Dec;36(6):462-477 [FREE Full text] [doi: 10.1016/j.jbi.2003.11.003] [Medline: 14759819]
8. Zhao Z, Yang Z, Luo L, Lin H, Wang J. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. Bioinformatics 2016 Dec 15;32(22):3444-3453 [FREE Full text] [doi: 10.1093/bioinformatics/btw486] [Medline: 27466626]
9. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Mar 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
10. Yao Y, Ye D, Li P, Han X, Lin Y, Liu Z, et al. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019 Presented at: 57th Annual Meeting of the Association for Computational Linguistics; July 2019; Florence, Italy [doi: 10.18653/v1/P19-1074]
11. Jung S, Lee T, Cheng C, Buble K, Zheng P, Yu J, et al. 15 years of GDR: New data and functionality in the Genome Database for Rosaceae. Nucleic Acids Res 2019 Jan 08;47(D1):D1137-D1145 [FREE Full text] [doi: 10.1093/nar/gky1000] [Medline: 30357347]
12. Xing R, Luo J, Song T. BioRel: towards large-scale biomedical relation extraction. BMC Bioinformatics 2020 Dec 16;21(Suppl 16):543 [FREE Full text] [doi: 10.1186/s12859-020-03889-5] [Medline: 33323106]
13. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 01;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
14. CT-PubMedBERT-RE. Hugging Face Hub. URL: https://huggingface.co/zhangzeyu/CT-PubMedBERT-RE [accessed 2023-08-31]
15. CT-PubMedBERT-RE-fine-tuned-type. Hugging Face Hub. URL: https://huggingface.co/zhangzeyu/CT-PubMedBERT-RE-fine-tuned-type [accessed 2023-08-31]
16. CT-PubMedBERT-RE-fine-tuned-typecode. Hugging Face Hub. URL: https://huggingface.co/zhangzeyu/CT-PubMedBERT-RE-fine-tuned-typecode [accessed 2023-08-31]
17. CT-PubMedBERT-RE-fine-tuned-group. Hugging Face Hub. URL: https://huggingface.co/zhangzeyu/CT-PubMedBERT-RE-fine-tuned-group [accessed 2023-08-31]
18. CT-PubMedBERT-RE-fine-tuned-groupabb. Hugging Face Hub. URL: https://huggingface.co/zhangzeyu/CT-PubMedBERT-RE-fine-tuned-groupabb [accessed 2023-08-31]
19. BIORE (RE tool based on CT_PubMedBERT). Docker Hub. URL: https://hub.docker.com/r/zhangzeyu/biore [accessed 2023-08-31]
20. Chen Q, Allot A, Lu Z. LitCovid: an open database of COVID-19 literature. Nucleic Acids Res 2021 Jan 08;49(D1):D1534-D1540 [FREE Full text] [doi: 10.1093/nar/gkaa952] [Medline: 33166392]
21. Lever J, Altman RB. Analyzing the vast coronavirus literature with CoronaCentral. Proc Natl Acad Sci U S A 2021 Jul 08;118(23):e2100766118 [FREE Full text] [doi: 10.1073/pnas.2100766118] [Medline: 34016708]
22. Wei C, Kao H, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res 2013 Jul;41(Web Server issue):W518-W522 [FREE Full text] [doi: 10.1093/nar/gkt441] [Medline: 23703206]
23. Wei C, Allot A, Leaman R, Lu Z. PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res 2019 Jul 02;47(W1):W587-W593 [doi: 10.1093/nar/gkz389] [Medline: 31114887]
24. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv. 2019. URL: https://arxiv.org/abs/1909.11942 [accessed 2023-08-31]
25. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. 2019. URL: https://arxiv.org/abs/1810.04805 [accessed 2023-08-31]
26. Peters M, Neumann M, Zettlemoyer L, Yih W. Dissecting Contextual Word Embeddings: Architecture and Representation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018 Presented at: Conference on Empirical Methods in Natural Language Processing; October-November 2018; Brussels, Belgium [doi: 10.18653/v1/D18-1179]
27. He P, Liu X, Gao J, Chen W. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv. 2020. URL: https://arxiv.org/abs/2006.03654 [accessed 2023-08-31]
28. Clark K, Luong M, Le Q, Manning C. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. arXiv. 2020. URL: https://arxiv.org/abs/2003.10555 [accessed 2023-08-31]

29. Sun Y, Wang S, Li Y, Feng S, Tian H, Wu H, et al. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2020 Presented at: Thirty-Fourth AAAI Conference on Artificial Intelligence; February 7–12, 2020; New York, NY p. 8968-8975 [doi: 10.1609/aaai.v34i05.6428]
30. Xu Y, Li M, Cui L, Huang S, Wei F, Zhou M. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. arXiv. 2019. URL: https://arxiv.org/abs/1912.13318 [accessed 2023-08-31]
31. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv. 2019. URL: https://arxiv.org/abs/1907.11692 [accessed 2023-08-31]
32. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le Q. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv. 2019. URL: https://arxiv.org/abs/1906.08237 [accessed 2023-08-31]
33. Alrowili S, Shanker V. BioM-Transformers: Building Large Biomedical Language Models with BERT, ALBERT and ELECTRA. In: Proceedings of the 20th Workshop on Biomedical Language Processing. 2021 Presented at: 20th Workshop on Biomedical Language Processing; June 2021; Online [doi: 10.18653/v1/2021.bionlp-1.24]
34. Tinn R, Cheng H, Gu Y, Usuyama N, Liu X, Naumann T, et al. Fine-Tuning Large Neural Language Models for Biomedical Natural Language Processing. arXiv. 2021. URL: https://arxiv.org/abs/2112.07869 [accessed 2023-08-31]
35. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, et al. Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020 Presented at: 58th Annual Meeting of the Association for Computational Linguistics; July 2020; Online [doi: 10.18653/v1/2020.acl-main.740]
36. roberta-pubmed (Roberta-Base fine-tuned on PubMed Abstract). Hugging Face. URL: https://huggingface.co/raynardj/roberta-pubmed [accessed 2023-08-31]
37. Beltagy I, Lo K, Cohan A. SciBERT: A Pretrained Language Model for Scientific Text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019 Presented at: Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); November 2019; Hong Kong, China [doi: 10.18653/v1/D19-1371]
38. Alsentzer E, Murphy J, Boag W, Weng W, Jindi D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019 Presented at: 2nd Clinical Natural Language Processing Workshop; June 2019; Minneapolis, MN [doi: 10.18653/v1/W19-1909]
39. Peng Y, Yan S, Lu Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In: Proceedings of the 18th BioNLP Workshop and Shared Task. 2019 Presented at: 18th BioNLP Workshop and Shared Task; August 2019; Florence, Italy [doi: 10.18653/v1/W19-5006]
40. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Trans. Comput. Healthcare 2021 Oct 15;3(1):1-23 [doi: 10.1145/3458754]
41. Wu S, He Y. Enriching Pre-trained Language Model with Entity Information for Relation Classification. arXiv. 2019. URL: https://arxiv.org/abs/1905.08284 [accessed 2023-08-31]
42. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-Art Natural Language Processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020 Presented at: Conference on Empirical Methods in Natural Language Processing: System Demonstrations; October 2020; Online [doi: 10.18653/v1/2020.emnlp-demos.6]
43. Sosa DN, Derry A, Guo M, Wei E, Brinton C, Altman RB. A Literature-Based Knowledge Graph Embedding Method for Identifying Drug Repurposing Opportunities in Rare Diseases. Pac Symp Biocomput 2020;25:463-474 [FREE Full text] [Medline: 31797619]
44. Percha B, Altman RB. A global network of biomedical relationships derived from text. Bioinformatics 2018 Aug 01;34(15):2614-2624 [FREE Full text] [doi: 10.1093/bioinformatics/bty114] [Medline: 29490008]
45. Coşkun M, Koyutürk M. Node similarity-based graph convolution for link prediction in biological networks. Bioinformatics 2021 Dec 07;37(23):4501-4508 [FREE Full text] [doi: 10.1093/bioinformatics/btab464] [Medline: 34152393]
46. Li G, Hilgenfeld R, Whitley R, De Clercq E. Therapeutic strategies for COVID-19: progress and lessons learned. Nat Rev Drug Discov 2023 Jul;22(6):449-475 [FREE Full text] [doi: 10.1038/s41573-023-00672-y] [Medline: 37076602]
47. Ho JSY, Mok BW, Campisi L, Jordan T, Yildiz S, Parameswaran S, et al. TOP1 inhibition therapy protects against SARS-CoV-2-induced lethal inflammation. Cell 2021 May 13;184(10):2618-2632.e17 [FREE Full text] [doi: 10.1016/j.cell.2021.03.051] [Medline: 33836156]
48. Neo4j COVID-19 relation graph database. bmtongji. URL: http://bmtongji.cn/redata/neo4j-covid-relation-database.dump [accessed 2023-08-31]
49. Rabiu Abubakar A, Ahmad R, Rowaiye AB, Rahman S, Iskandar K, Dutta S, et al. Targeting Specific Checkpoints in the Management of SARS-CoV-2 Induced Cytokine Storm. Life (Basel) 2022 Mar 25;12(4):478 [FREE Full text] [doi: 10.3390/life12040478] [Medline: 35454970]
50. Kaye AG, Siegel R. The efficacy of IL-6 inhibitor Tocilizumab in reducing severe COVID-19 mortality: a systematic review. PeerJ 2020;8:e10322 [FREE Full text] [doi: 10.7717/peerj.10322] [Medline: 33194450]

51. Zheng K, Xu Y, Guo Y, Diao L, Kong X, Wan X, et al. Efficacy and safety of tocilizumab in COVID-19 patients. Aging (Albany NY) 2020 Oct 08;12(19):18878-18888 [FREE Full text] [doi: 10.18632/aging.103988] [Medline: 33031060]
52. Zayed S, Belal F. Determination of the Monoclonal Antibody Tocilizumab by a Validated Micellar Electrokinetic Chromatography Method. Chromatographia 2022;85(5):481-488 [FREE Full text] [doi: 10.1007/s10337-022-04148-w] [Medline: 35382455]
53. Xu X, Han M, Li T, Sun W, Wang D, Fu B, et al. Effective treatment of severe COVID-19 patients with tocilizumab. Proc Natl Acad Sci U S A 2020 May 19;117(20):10970-10975 [FREE Full text] [doi: 10.1073/pnas.2005615117] [Medline: 32350134]
54. Novel drug path data. bmtongji. URL: http://bmtongji.cn/redata/novel-drug-path-data.zip [accessed 2023-08-31]
55. Donohue JF, Mahler DA, Sethi S. Revefenacin: A Once-Daily, Long-Acting Bronchodilator For Nebulized Treatment Of COPD. Int J Chron Obstruct Pulmon Dis 2019;14:2947-2958 [doi: 10.2147/COPD.S157654] [Medline: 31908443]
56. Singh D, Mathioudakis AG, Higham A. Chronic obstructive pulmonary disease and COVID-19: interrelationships. Curr Opin Pulm Med 2022 Mar 01;28(2):76-83 [FREE Full text] [doi: 10.1097/MCP.0000000000000834] [Medline: 34690257]
57. Beyerstedt S, Casaro EB, Rangel EB. COVID-19: angiotensin-converting enzyme 2 (ACE2) expression and tissue susceptibility to SARS-CoV-2 infection. Eur J Clin Microbiol Infect Dis 2021 May;40(5):905-919 [FREE Full text] [doi: 10.1007/s10096-020-04138-6] [Medline: 33389262]
58. Leung JM, Yang CX, Tam A, Shaipanich T, Hackett T, Singhera GK, et al. ACE-2 expression in the small airway epithelia of smokers and COPD patients: implications for COVID-19. Eur Respir J 2020 May 08;55(5):2000688 [FREE Full text] [doi: 10.1183/13993003.00688-2020] [Medline: 32269089]
59. Djokovic N, Ruzic D, Djikic T, Cvijic S, Ignjatovic J, Ibric S, et al. An Integrative in silico Drug Repurposing Approach for Identification of Potential Inhibitors of SARS-CoV-2 Main Protease. Mol Inform 2021 May;40(5):e2000187 [FREE Full text] [doi: 10.1002/minf.202000187] [Medline: 33787066]
60. Xia XM, Wang H. Gastroesophageal Reflux Disease Relief in Patients Treated with Rabeprazole 20 mg versus Omeprazole 20 mg: A Meta-Analysis. Gastroenterol Res Pract 2013;2013:327571 [FREE Full text] [doi: 10.1155/2013/327571] [Medline: 24106498]
61. Pace F, Annese V, Prada A, Zambelli A, Casalini S, Nardini P, Italian Rabeprazole Study Group. Rabeprazole is equivalent to omeprazole in the treatment of erosive gastro-oesophageal reflux disease. A randomised, double-blind, comparative study of rabeprazole and omeprazole 20 mg in acute treatment of reflux oesophagitis, followed by a maintenance open-label, low-dose therapy with rabeprazole. Dig Liver Dis 2005 Oct;37(10):741-750 [doi: 10.1016/j.dld.2005.04.026] [Medline: 16024305]
62. Hussain A, Naughton DP, Barker J. Potential Effects of Ibuprofen, Remdesivir and Omeprazole on Dexamethasone Metabolism in Control Sprague Dawley Male Rat Liver Microsomes (Drugs Often Used Together Alongside COVID-19 Treatment). Molecules 2022 Mar 30;27(7):2238 [FREE Full text] [doi: 10.3390/molecules27072238] [Medline: 35408639]
63. Ray A, Sharma S, Sadasivam B. The Potential Therapeutic Role of Proton Pump Inhibitors in COVID-19: Hypotheses Based on Existing Evidences. Drug Res (Stuttg) 2020 Oct;70(10):484-488 [FREE Full text] [doi: 10.1055/a-1236-3041] [Medline: 32877948]
64. Birkenkamp KU, Tuyt LM, Lummen C, Wierenga AT, Kruijer W, Vellenga E. The p38 MAP kinase inhibitor SB203580 enhances nuclear factor-kappa B transcriptional activity by a non-specific effect upon the ERK pathway. Br J Pharmacol 2000 Oct;131(1):99-107 [FREE Full text] [doi: 10.1038/sj.bjp.0703534] [Medline: 10960075]
65. Mohd Zawawi Z, Kalyanasundram J, Mohd Zain R, Thayan R, Basri DF, Yap WB. Prospective Roles of Tumor Necrosis Factor-Alpha (TNF-α) in COVID-19: Prognosis, Therapeutic and Management. Int J Mol Sci 2023 Mar 24;24(7):6142 [FREE Full text] [doi: 10.3390/ijms24076142] [Medline: 37047115]
66. Coomes EA, Haghbayan H. Interleukin-6 in Covid-19: A systematic review and meta-analysis. Rev Med Virol 2020 Dec;30(6):1-9 [FREE Full text] [doi: 10.1002/rmv.2141] [Medline: 32845568]
67. Grimes JM, Grimes KV. p38 MAPK inhibition: A promising therapeutic approach for COVID-19. J Mol Cell Cardiol 2020 Jul;144:63-65 [FREE Full text] [doi: 10.1016/j.yjmcc.2020.05.007] [Medline: 32422320]
68. Faist A, Schloer S, Mecate-Zambrano A, Janowski J, Schreiber A, Boergeling Y, et al. Inhibition of p38 signaling curtails the SARS-CoV-2 induced inflammatory response but retains the IFN-dependent antiviral defense of the lung epithelial barrier. Antiviral Res 2023 Jan;209:105475 [FREE Full text] [doi: 10.1016/j.antiviral.2022.105475] [Medline: 36423831]
69. Malekinejad Z, Baghbanzadeh A, Nakhlband A, Baradaran B, Jafari S, Bagheri Y, et al. Recent clinical findings on the role of kinase inhibitors in COVID-19 management. Life Sci 2022 Oct 01;306:120809 [FREE Full text] [doi: 10.1016/j.lfs.2022.120809] [Medline: 35841979]
70. CoronaCentral data. bmtongji. URL: http://bmtongji.cn/redata/coronacentral-data.xlsx [accessed 2023-08-31]
71. Pérez-Pérez M, Rabal O, Pérez-Rodríguez G, Vazquez M, Fdez-Riverola F, Oyarzabal J, et al. Evaluation of chemical and gene/protein entity recognition systems at BioCreative V.5: the CEMP and GPRO patents tracks. In: Proceedings of the BioCreative V.5 Challenge Evaluation Workshop. 2017 Presented at: BioCreative V.5 Challenge Evaluation Workshop; April 25-27, 2017; Barcelona, Spain

72. Krallinger M, Leitner F, Rabal O, Vazquez M, Oyarzabal J, Valencia A. CHEMDNER: The drugs and chemical names extraction challenge. J Cheminform 2015;7(Suppl 1 Text mining for chemistry and the CHEMDNER track):S1 [FREE Full text] [doi: 10.1186/1758-2946-7-S1-S1] [Medline: 25810766]

73. Herrero-Zazo M, Segura-Bedmar I, Martínez P, Declerck T. The DDI corpus: an annotated corpus with pharmacological substances and drug-drug interactions. J Biomed Inform 2013 Oct;46(5):914-920 [FREE Full text] [doi: 10.1016/j.jbi.2013.07.011] [Medline: 23906817]

74. Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. Nat Genet 2004 May;36(5):431-432 [doi: 10.1038/ng0504-431] [Medline: 15118671]

75. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. arXiv. 2013. URL: https://arxiv.org/abs/1301.3781 [accessed 2023-08-31]

76. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality. arXiv. 2013. URL: https://arxiv.org/abs/1310.4546 [accessed 2023-08-31]

77. Singh B, Ryan H, Kredo T, Chaplin M, Fletcher T. Chloroquine or hydroxychloroquine for prevention and treatment of COVID-19. Cochrane Database Syst Rev 2021 Mar 12;2(2):CD013587 [FREE Full text] [doi: 10.1002/14651858.CD013587.pub2] [Medline: 33624299]

78. Sogut O, Can MM, Guven R, Kaplan O, Ergenc H, Umit TB, et al. Safety and efficacy of hydroxychloroquine in 152 outpatients with confirmed COVID-19: A pilot observational study. Am J Emerg Med 2021 Mar;40:41-46 [FREE Full text] [doi: 10.1016/j.ajem.2020.12.014] [Medline: 33348222]

79. Mohana A, Sulaiman T, Mahmoud N, Hassanein M, Alfaifi A, Alenazi E, et al. Hydroxychloroquine Safety Outcome within Approved Therapeutic Protocol for COVID-19 Outpatients in Saudi Arabia. Int J Infect Dis 2021 Jan;102:110-114 [FREE Full text] [doi: 10.1016/j.ijid.2020.10.031] [Medline: 33075525]

80. Wang M, Cao R, Zhang L, Yang X, Liu J, Xu M, et al. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. Cell Res 2020 Mar;30(3):269-271 [FREE Full text] [doi: 10.1038/s41422-020-0282-0] [Medline: 32020029]

81. Looi M. What are the latest covid drugs and treatments? BMJ 2023 May 03;381:872 [doi: 10.1136/bmj.p872] [Medline: 37137505]

82. Wang B, Xie Q, Pei J, Chen Z, Tiwari P, Li Z, et al. Pre-trained Language Models in Biomedical Domain: A Systematic Survey. ACM Comput. Surv 2023 Aug:3611651 [doi: 10.1145/3611651]

83. Wei C, Peng Y, Leaman R, Davis AP, Mattingly CJ, Li J, et al. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. Database (Oxford) 2016;2016:baw032 [FREE Full text] [doi: 10.1093/database/baw032] [Medline: 26994911]

84. Zhang Y, Zheng W, Lin H, Wang J, Yang Z, Dumontier M. Drug-drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths. Bioinformatics 2018 Mar 01;34(5):828-835 [FREE Full text] [doi: 10.1093/bioinformatics/btx659] [Medline: 29077847]

Abbreviations

COPD: chronic obstructive pulmonary disease
CUI: Concept Unique Identifier
IL-6: interleukin-6
KG: knowledge graph
MLM: masked language modeling
NLP: natural language processing
PCA: principal component analysis
RE: relation extraction
RRF: rich release format
TNFα: tumor necrosis factor-α
UMLS: Unified Medical Language System

Edited by A Mavragani; submitted 11.04.23; peer-reviewed by R Leaman, X Li; comments to author 09.05.23; revised version received 03.07.23; accepted 25.08.23; published 20.09.23

Please cite as:
Zhang Z, Fang M, Wu R, Zong H, Huang H, Tong Y, Xie Y, Cheng S, Wei Z, Crabbe MJC, Zhang X, Wang Y
Large-Scale Biomedical Relation Extraction Across Diverse Relation Types: Model Development and Usability Study on COVID-19
J Med Internet Res 2023;25:e48115
URL: https://www.jmir.org/2023/1/e48115
doi: 10.2196/48115
PMID: 37632414