Mining Microbe-disease Interactions From Literature via a Transfer Learning Model

Chengkun Wu (chengkun_wu@nudt.edu.cn)
State Key Laboratory of High-Performance Computing, National University of Defense Technology
Xinyi Xiao
College of Computer, National University of Defense Technology
Canqun Yang
College of Computer, National University of Defense Technology
JinXiang Chen
Department of General Surgery, Xiangya Hospital, Central South University
Jiacai Yi
College of Computer, National University of Defense Technology
Yanlong Qiu
College of Computer, National University of Defense Technology

Research Article

Keywords: Microbe-Disease Interactions, Named-entity recognition, Relation extraction, Transfer Learning

Posted Date: June 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-558906/v1

RESEARCH

Mining microbe-disease interactions from literature via a transfer learning model

Chengkun Wu[1,2*], Xinyi Xiao[2], Canqun Yang[2], JinXiang Chen[3], Jiacai Yi[2] and Yanlong Qiu[2]

*Correspondence:
chengkun_wu@nudt.edu.cn
[1]State Key Laboratory of High-Performance Computing, National University of Defense Technology, 410073, Changsha, China
[2]College of Computer, National University of Defense Technology, 410073, Changsha, China
[3]Department of General Surgery, Xiangya Hospital, Central South University, 410008, Changsha, China
Full list of author information is available at the end of the article

Abstract

Background: Interactions of microbes and diseases are of great importance for biomedical research. However, large-scale curated databases for microbe-disease interactions are missing, as the amount of related literature is enormous and the curation process is costly and time-consuming. In this paper, we aim to construct a large-scale database for microbe-disease interactions automatically. We attained this goal via applying text mining methods based on a deep learning model with a moderate curation cost. We also built a user-friendly web interface to allow researchers navigate and query desired information.

Results: For curation, we manually constructed a golden-standard corpora (GSC) and a sliver-standard corpora (SSC) for microbe-disease interactions. Then we proposed a text mining framework for microbe-disease interaction extraction without having to build a model from scratch. Firstly, we applied named entity recognition (NER) tools to detect microbe and disease mentions from texts. Then we transferred a deep learning model BERE to recognize relations between entities, which was originally built for drug-target interactions or drug-drug interactions. The introduction of SSC for model fine-tuning greatly improves the performance of detection for microbe-disease interactions, with an average reduction in error of approximately 10%. The resulting MDIDB website offers data browsing, custom search for specific diseases or microbes as well as batch download.

Conclusions: Evaluation results demonstrate that our method outperform the baseline model (rule-based PKDE4J) with an average F_1-score of 73.81%. For further validation, we randomly sampled nearly 1,000 predicted interactions by our model, and manually checked the correctness of each interaction, which gives a 73% accuracy. The MDIDB webiste is freely avaliable throuth http://dbmdi.com/index/

Keywords: Microbe-Disease Interactions; Named-entity recognition; Relation extraction; Transfer Learning

Background

Microbiota in the human body is of great significance to human health. Pathogenic microorganisms are the chief culprit for many human diseases [1], such as the SARS outbreak in 2003 [2, 3] and the avian influenza (HPAI) [4] in the past few years, as well as inflammatory bowel disease (IBD) caused by enteric human virome [5, 6]. Studies have even shown that there is a close connection between mental illness and gut microbes [7, 8]. Through the detection of gut microbes in patients with chronic heart failure(CHF), compared with normal individuals, patients with heart failure had higher levels of gram-negative bacteria and *Candida* in the intestine, and

intestinal permeability was increased, promoting the process of CHF [9]. The gut flora can also impact arthritis(AR), the work in [10] applied 16S rDNA sequencing to sequence the gut microbiota of patients and healthy individuals, and found that the abundance of the gut microbiota reduced significantly in patients with AR. At the same time, the extent of the reduction was positively correlated with the length of the disease course as well as with the severity of the disease. Therefore, it is essential to be able to efficiently explore relations between microbes and diseases, which is currently not feasible due to the fact that most information is buried in the vast amount of unstructured biomedical literature.

The first human microbial-disease association database (HMDAD) was established to provide experimental data for microbial disease association research. The database only contains 39 disease entities and 292 microbial species, and the relationship between the two entities is established at the text level [11]. Most studies on the prediction of microbial disease associations are based on this database like KATZHMDA [12], NCPHMDA [13], MDLPHMDA [14], RNMFMDA [15]. However, due to the limited types of diseases and microorganisms included in this database, a large amount of information in biomedical texts has not been fully mined. MicroPhenoDB is a recent work of the relationships between disease phenotypes, pathogenic microbes, and core genes. It was built by a manual review process and a calculation method, which collects the IDSA guideline data, the manual curate data resource, and traceable literature with different weights to calculate the score between microbes and diseases [16]. Most studies on the relationship between microorganisms and diseases need many human resources. [17] proposed a hierarchical extended short-term memory network and an ensemble parser model to capture the trigger word of each disease-microbe entity pair in a single instance by firstly utilize binary classification to judge whether are some relations between two entities and then catch the relation word of them. PubMed is a free database for biomedical and life sciences literature, with over 70 million abstracts and more than 7 million full-text articles. By March 2021, 64510 records were retrieved from PubMed and 64259 full-text records were retrieved from PMC by the 'microbe' query. As illustrated in Figure 1, the amount of microbe-related literature is increasing rapidly in the recent twenty years, which makes it difficult for microbe researchers to identify, retrieve and assimilate all relevant publications. Hence, automated text mining is an essential tool to discover the valuable information hidden in this enormous amount of literature.

Biomedical named entity recognition(NER) is a fundamental task for understanding biomedical literature, which is mostly presented as non-structural texts injected with a large number of specialized terms. A number of successful NER tools have been developed for diseases [18], genes/proteins [19, 20] , species [21], chemicals [22], etc. In this work, we use DNORM [18] to recognize disease entities, which is a machine learning based toolkit for disease NER and normalization. For microbes, there is no such tool available and we have to build our own method. Biomedical relation extraction(BioRE) aims to automatically capture relations between two entities from NER results. The entity relationship not only facilitates the acquisition of domain knowledge by researchers in the biomedical field, but also enables the automated processing of biomedical information, and promotes the development

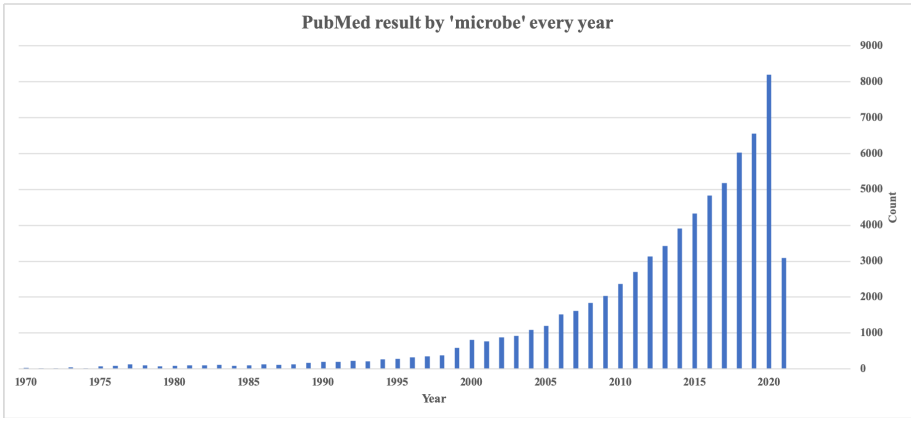


Figure 1: Number of PubMed articles returned by the 'microbe' query.

of research tools in the biomedical field and the development of information in the medical field. Previous studies and datasets on BioRE already discussed about protein-protein interactions(PPIs) [23], drug-drug interactions(DDI), drug-target interactions(DTIs), etc. Still, the classification of the relation between microbe and disease has no clear definition.

Machine learning and deep learning methods rely heavily on manually labeled data sets, and human annotation is costly and time-consuming. Transfer learning has been successfully utilized in many natural language processing fields such as text classification [24], named entity recognition [25]. It extracts knowledge from one or more source domains and applies it to the target domain. [25] applied this idea on biomedical named entity recognition, a deep neural network(DNN) was trained on large silver-standard corpora with noise and then transferred to small gold-standard corpora. It indeed showed a significant improvement on 23 gold-standard corpora covering chemicals, disease, species, and genes/proteins. Inspired by the work of transfer learning for biomedical named entity recognition [25], we introduced transfer learning to address our problem of microbe-disease interaction text mining from literature.

Our main contributions can be summarized as follows: (1) we utilized NER tools to locate microbe and disease entities from a large collection of related literature; (2) we manually created two microbe-disease interaction corpora for the following training process, including a gold-standard and a silver-standard; (3) we applied transfer Learning to perform microbe-disease relation extraction without the need for a large-scale curation; (4) we developed a user-friendly website to help biomedical researchers find valuable information about diseases and microbes.

Methods

Data preparation

Literature data used in this work was download from PMC (https://www.ncbi.nlm.nih.gov/pmc) and PebMed (https://pubmed.ncbi.nlm.nih.gov), by searching keyword "microbe", a list of PubMed IDs can be got (accessed on March, 2021). We used Aspera (https://www.ibm.com/products/aspera) as a tool to download the PubMed database on NCBI, then retrieved abstracts according to listed

PubMed IDs. If the corresponding full-text is available in PMC, we then use Eutils, a tool provided by PMC, to obtain the XML file of the full-text. A total collection of 24,256 articles was built as our data sources. To locate microbe mentions in texts, we built a specialized dictionary of microbe names collected from Human Microbe-Disease Association Database [11] (HMDAD, http://www.cuilab.cn/hmdad), Virtual Metabolic Human [26] (VMH, https://vmh.life) and Disbiome [27] (https://disbiome.ugent.be). The final microbe dictionary in included 3,775 microbes. Next, we retrieved the taxonomy id of each microbe name to prepare for the BioNER procedure. Figure 2 shows the whole workflow of data preparation.

Named entities recognition(NER) and Relation extraction(RE)

In this study, we only considered the microbe-disease relation at the sentence level, so sentence splitting is necessary for pre-processing, which is carried out with NLTK, a Python natural language toolkit.

There is no readily available NER tool for microbes. LINNAEUS is a dictionary-based species name identification system for biomedical literature, performs with 94% recall and 97% precision at the mention level [21]. Using LINNAEUS and the microbe dictionary, we can track the microbial entities in the texts with the information of each entity's start and end position, which will be used as input data in the RE step(shown in Figure 2b). DNORM is a well-established disease name normalization model with a 0.782 micro-averaged F-measure and 0.809 macro-averaged F-measure performance. Normalized disease mentions are identified with their MeSH ids. An example of DNorm result is shown in Figure 2c.

A successful RE requires at least one microbe mention and one disease mention in the input sentence. The sentence instance will be in the format like Figure 2d, which is the input format of PKDE4J.

Once the sentences are correctly formatted, we removed those instances with more than 64 words as many longer sentences can lead to detection errors. We use a highly flexible and extensible relation extraction tool, PKDE4J, as the baseline method. It applies dependency tree-based rules to extract relationships among entities in sentences with two or more entities [28]. PKDE4J is based on dependency parsing technologies, which define rules to find the syntactic and grammatical structures and trigger words from sentences. Figure 2e shows the output format of PKDE4J. We got 96,670 instances after the relation extraction of PKDE4J. We also used PKDE4J to generate the silver-standard corpora(SSCs)(shown in Figure 2f).

Data Curation

Human annotated gold-standard corpora(GSCs)

To better evaluate the performance of our method, we curated gold-standard corpora with hand-labeled annotations. We employed PubTator Central (PTC, https://www.ncbi.nlm.nih.gov/research/pubtator/), a web-based system for automatic annotations of biomedical concepts in PubMed abstracts and PMC full-text texts, to mark entities with their MeSH ids and Taxonomy ids. Microbe-disease relation types are defined as follows:

- **positive:** This type is used to annotate microbe-disease entity pairs with a positive correlation, such as microbe will cause or aggravate disease, microbe will increase when disease occurs.

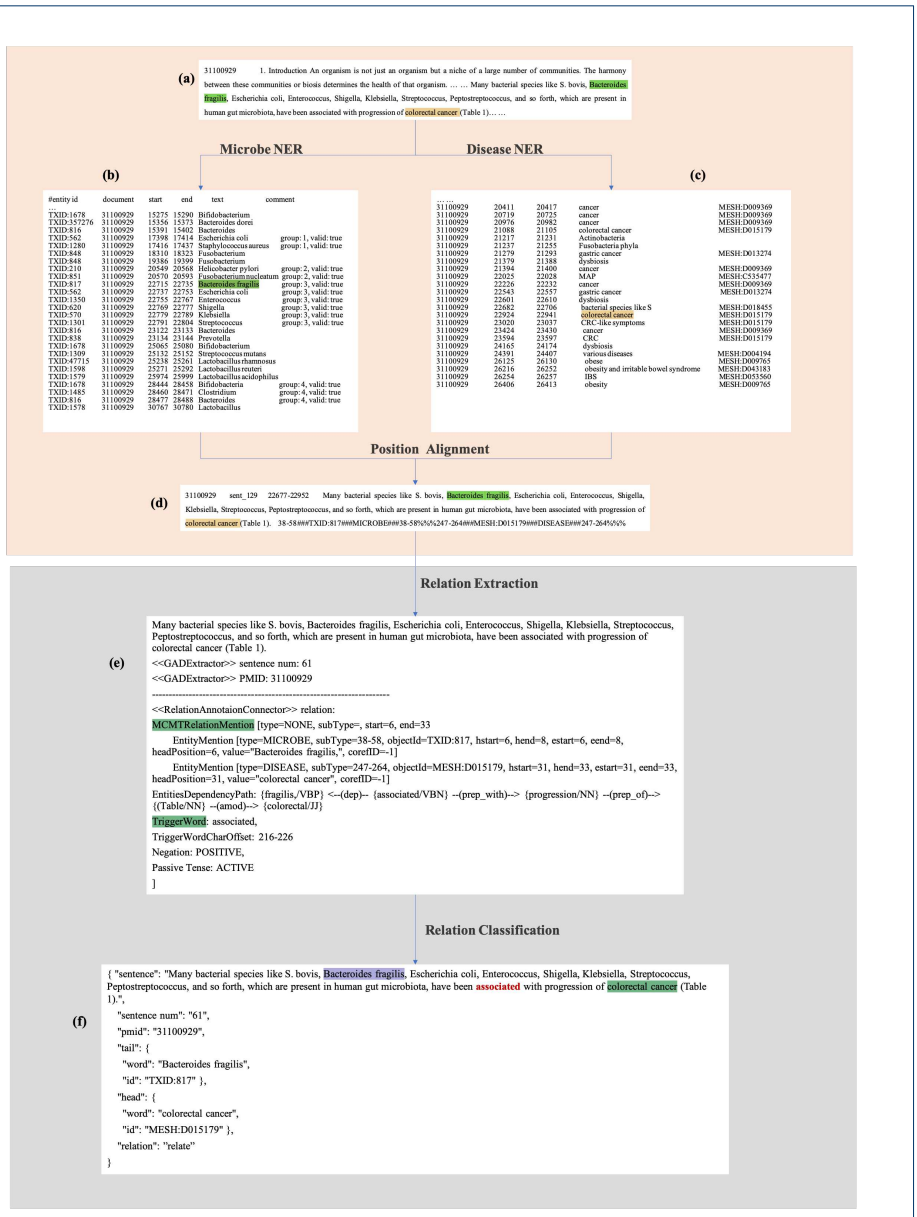
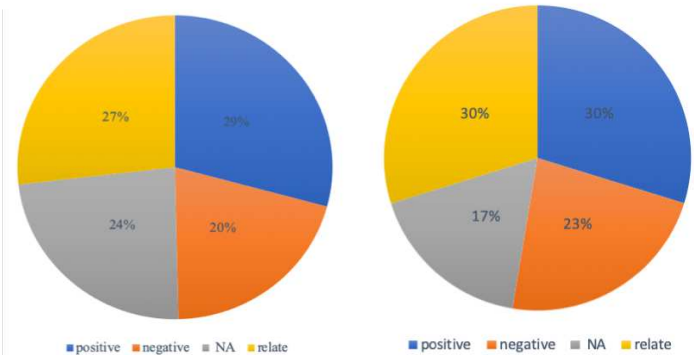


Figure 2: The workflow of data preparation. (a)The initial full-text data (b)The result of applying LINNAEUS to recognize microbe entities (c)The result of applying DNORM to recognize disease entities (d)The result of splitting texts into sentences and aligning the position of disease and microbe (e)The result of relation extraction using PKDE4J (f)The result of classification instance by the result of PKDE4J

- **negative:** This type is used to annotate microbe disease entity pairs that have a negative correlation, such as microbe can be a treatment for a disease, or microbe will decrease when disease occurs.



(a) The distribution for GSC (b) The distribution for SSC

Figure 3: The distribution of relation types for GSC and SSC

- **relate:** This type is used when a microbe disease entity pair appears in the instance and described they are related with each other without additional information
- **NA:** This type is used when a microbe disease entity pair appears in the instance, but the relation of these two entities, which has not been described yet.

We removed the instance if the instance has no tag or has a wrong tag in PTC. Then the instances were classified into the above four relations we defined. Finally, we got a set of 1,100 manually annotated instances, and we use it as the gold-standard corpora for transfer learning and performance evaluation.

Silver-standard Corpora(SSCs)

The cost of enlarging the size of GSC is very high as each sample needs to be carefully reviewed. Due to the high cost, the size of the GSC is very limited. To provide more training samples, we built a silver-standard corpora with automated tools rather than human annotation. This means SSCs might contain many incorrect annotations (noise). To do this, we applied PKDE4J on over 20,000 articles related to microbe. The results of PKDE4J include information on the relation between microbe and disease, 'RelationMentionType' and 'Trigger words' (shows in Figure 2e), which can be used as auxiliary information for relation type annotation. For example, if one instance is tagged with the RelationMentionType 'increased', we assign the instance with a relation type 'positive'. Results with RelationMentionType "JUXTAPOSE" were removed. The 'Trigger word' tag was also utilized to define the relation type.

At last, each instance will be classified in one relation type in positive, negative, relate, NA. Instance appeared in the GSC were removed from SSC. The resulting SSC dataset contains 12,959 samples, and it is used as a major training data source for the transfer learning procedure. Figure 3 shows distribution of relation types for both GSC and SSC.

Transfer learning with BERE

Most machine learning application scenarios requires a lot of labeled data for supervised learning. However, annotating data is a tedious and costly task. We address this problem via transfer learning. BERE is a deep learning framework to extract drug related relations from literature automatically. This model uses latent tree learning and self-attention techniques to capture the syntactic information of the sentence. The input sentences firstly translate in the vector representations of words. Pre-trained word embedding are from http://bio.nlplab.org/. Each word in sentence will be represented in a concatenation of a 200 dimensions word embedding and a randomly initialized 50 dimensions POS embedding. Then Bi-GRU and self-attention mechanisms are applied to encode short and long-range dependencies between words. Gumbel Tree-GRU can implicitly learn the syntactic features of sentences. And it embeds the contextual features of two entities into the sentence representation. Lastly, a classifier will predict the relation between two entities. It shows great performance on the relation between drug-drug interaction, and the authors applied the model on a distantly supervised drug-target interaction dataset. A detailed description of BERE's architecture is explained in [29].

In the study of BERE, they use the DDI'13 dataset to demonstrate the performance of their model, and it turns out that the BERE model is better than six other baseline methods on the DDI'13 dataset. They then construct a distantly supervised Drug-Target interaction(DTI) dataset, which inspired us to make use of BERE to build a disease-microbe interaction dataset. In this work, we used the INS mode of BERE, which predicts each sentence instance into an individual class.

Training and Evaluation metrics

To better verify the effectiveness of BERE on MDI dataset with transfer learning, we compared the performance of BERE_TL and BERE_g. The SSC datasets were split into three disjoint subsets, 12000 samples for training the model, and 1000 of those data as the validation set. The rest of the samples were used as a test set for the final evaluation. This split operation on SSC was applied twice to take the average result to reduce the prediction bias. And the GSC was randomly separated as: 800 for the train set, 100 for the valid set, and 200 for the test set.

To demonstrate the role of transfer learning, we conducted 5-fold cross-validation of the BERE_TL and BERE_g on GSC. The typical evaluation indicators Precision, Recall, and F_1-score were used as evaluation metrics. The precision rate calculates the correct classified samples in all model samples, and the recall rate calculates the proportion of correct predicted correct positive samples. F_1 is a measure of precision and recall. We also compute the average percent reduction in F_1-score as the same as [25]:

$$\frac{F_1^{TL} - F_1^{baseline}}{100 - F_1^{baseline}} * 100$$

Web implementation

The website of MDIDB is implemented in the framework of Django, with AJAX loading dynamic data from a database based on MySql. The visual front-end page

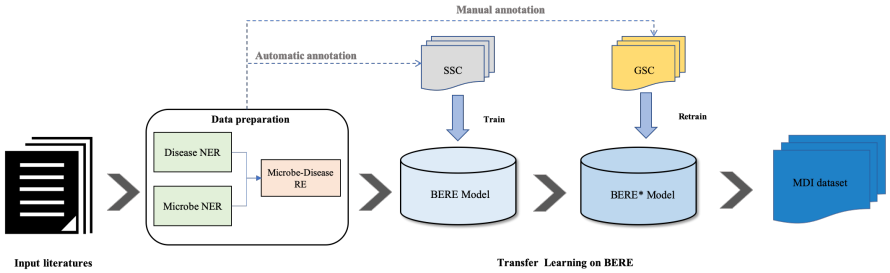


Figure 4: The procedures of mining information from literature

is built on the basis of Bootstrap 4, and the chart is based on the visual plug-in echart. Data access and operations were provided in a user-friendly way. By clicking the related term, users can browse the whole relevant microbe and disease list and the relevant statistical chart information of related word cloud chart and pie chart. Simultaneously, the website provides a search function for users to retrieve the information they are interested in. The relevant result data set of the paper can also be obtained from the download page.

The whole system is based on NLP algorithms for text mining of massive biological literature. Figure 4 shows the workflow of the whole text-mining system. After a series of post-processing, text mining results are stored in the database and operated by the backend server. Finally, we got a visual website containing 1,198 diseases, 1,065 microorganisms and 44,900 records of their relationship data.

Results

To prove that the BERE model can lay a solid foundation for the detection of microbe-disease relations, we compared the performance of BERE on several datasets with the rule-based baseline PKDE4J(MDI). Table 1 compares the micro-averaged performance metrics of each dataset. The learning rate was set to 0.0001, the dropout rate to 0.5.

BERE_g(MDI) is generated by fine-tuning the original BERE model only on the GSC training set. Results of BERE(DDI) and BERE(DTI) come from the origin BERE paper.

As of yet, it is not clear whether the introduction of transfer learning on BERE can improve the performance of MDI detection. Thus we evaluated the performance on the MDI dataset with two modes: BERE_TL(MDI) introduces transfer learning on the GSC training set while BERE_g(MDI) directly applied the original BERE model.

As Table 1 shows, we can see that compared with PKDE4J(MDI), BERE_g(MDI) achieves a higher score of precision, recall, and F_1-score on the same MDI dataset. Moreover, BERE_g(MDI) achieves a comparable performance with BERE(DDI) and BERE(DTI).

Quantifying the performance of transfer learning

To highlight the effect of transfer learning, we compared the performances with or without transfer learning. The experiment was performed under five-fold Cross-Validation and the final result was computed by average. Table2 lists the results

Table 1: Comparison of baseline performance on different datasets

	Precision	Recall	F_1-score
BERE(DDI)	76.8	71.3	73.9
BERE(DTI)	73.8	54.2	62.5
BERE_g(MDI)	68.8	71.4	70.1
PKDE4J(MDI)	55.3	41.3	47.3

for the BERE_g(MDI) against BERE_TL(MDI). It is evident that transfer learning significantly improved precision, recall and F_1-score. In addition, it brings an average reduction in error of 12% on GSC. Figure 5 shows the Precision-Recall curve of and the AUPRC result of BERE with transfer learning.

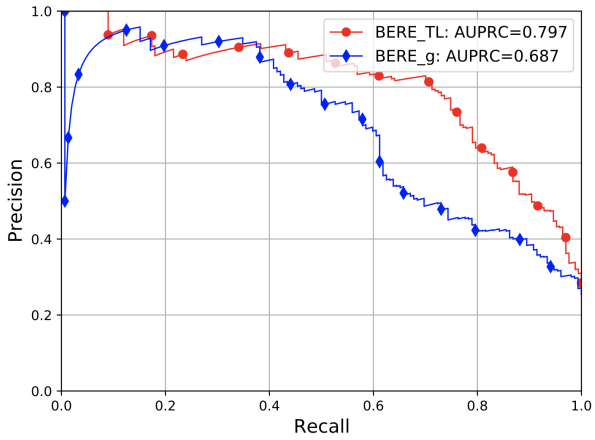


Figure 5: Comparisons of the Precision–Recall curves between BERE with or without transfer learning. The AUPRC and F_1-score for each method are on the top right contains.

Table 2: Results of 5-fold cross-validation

	BERE_TL(MDI)			BERE_base(MDI)		
	Precision(%)	Recall(%)	F_1-score(%)	Precision(%)	Recall(%)	F_1-score(%)
Fold-1	74.43	77.51	75.94	71.43	68.05	69.70
Fold-2	74.53	71.01	72.73	65.73	69.23	69.23
Fold-3	70.59	71.01	70.80	62.83	71.01	66.67
Fold-4	75.71	80.24	77.91	69.02	76.05	72.36
Fold-5	73.01	70.41	71.69	68.04	79.04	73.13
Average	**73.65**	**74.04**	**73.81**	67.41	72.68	70.22

Error analysis

We manually inspected some reported results of our model and we have the following observations:

Firstly, sentences with too many compound clauses may give rise errors. To improve this, we will need better NLP tools for semantic parsing or syntactic analysis of texts.

Secondly, some errors can be attributed to the NER tools. DNorm occasionally failed in cases of abbreviations and acronyms. For instance, 'WS' refers to wheat sensitivity in the article but DNorm tagged it as an abbreviation of disease 'Williams

Syndrome'. Pathologically related words can bring some misunderstanding too, 'syntrophic growth' was wrongly recognized as the disease 'Growth Disorders'. To reduce such errors, we will need better NER tools.

In addition, some texts might not even constitute a proper sentence. We noticed one example "Gastric cancer H. pylori, Porphyromonas, Neisseria, Prevotella pallens, Streptococcus sinensis, Lactobacillus coleohominis.." (PMID: 31236389), which was due to an improper representation of a table into text segments in the corresponding full-text XML document.

We randomly selected 1,000 predicted instances from the results by our model and manually checked each instance. 731 out of 1,000 were verified to be correct and 268 were verified to be wrongly predicted, which gives an accuracy of 73.1%. 914 instances were not found in the aforementioned database MicroPheno, but our manual inspection found that 633(69.2%) of them are correct and should be included.

To note, the recall of our method is around 74%, which means some useful information in literature might not be recovered. For instance, we know Bacillus cereus is a gram-positive bacteria that can produce toxin and causes diarrhea and we find some evidence by literature review [30, 31, 32]. However, this information was not included in our database. The reason is that our model only considers relation extraction at the sentence level. In some cases, useful information can only be mined across multiple sentences. We will leave that for future work. .

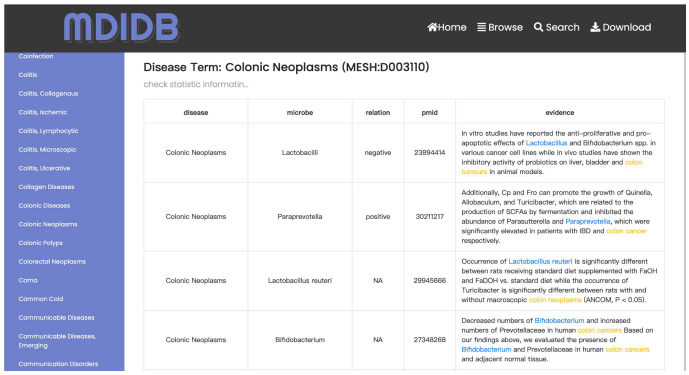
Searching on MDIDB website

In this section, we give examples on how to access MDIDB and retrieve useful information from our database.

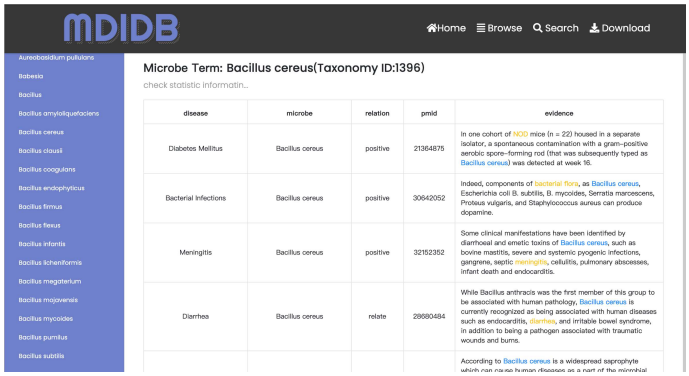
To demonstrate how to get related microbes by searching for disease names, we queried "Colonic Neoplasms", as illustrated in Figure 6(a). We obtained a list of microbe-disease relation records about colonial neoplasms, and each record has one evidence to support the classification of entity relation. The statistical chart result is shown in Figure 7(a), 7(b).

We can also search by microbe names. By searching microbe "Bacillus cereus"(Figure 6(b)), we got a list of related diseases, which includes Meningitis [33],Diabetes Mellitus [34],Dysentery, Endotoxemia [35],shown in(Figure 7(c), 7(d)).

MDIDB can generate top-ten pie charts for different queries and present an informative word-cloud for the most relevant microbes or diseases. For instance, the study [36] shows probiotics Lactobacilli can bring less abdominal discomfort for patients with colon cancer. [37] discussed the relations between Fusobacterium species and colon cancer. [38] had "protective" anti-cancer properties for colon cancer. Fusobacterium nucleatum is a gram-negative obligate anaerobic bacteria and can activate Wnt/beta-catenin signaling to accelerating proliferation of colon cancer cells [39, 40]. The relation of Clostridium and colon cancer was demonstrated in work [41], Clostridium is associated with progression of colonic cancer [42]. Moreover, [43] proves that Lachnospiraceae is linked to colon cancer, [44] found that Prevotella is associated with colon cancer, [45] suggested that Streptomyces can suppress colon tumorigenesis, [46] shows that some Streptococcus species are associated with colon



(a) Searching result of disease "Colonic Neoplasms"



(b) Searching result of microbe "Bacillus cereus"

Figure 6: Query results in MDIDB.

cancer. Colorectal Neoplasms and Colonic Neoplasms have a similar result of statistic chart, and as we know, colon cancer and colorectal cancer are equivalent in some literature.

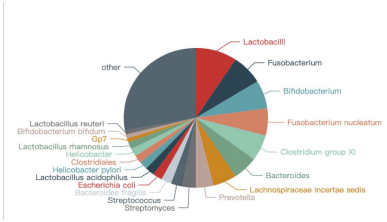
Discussion

Extracting information from a large number of scientific literature to assist relevant microbial researchers in their research can provide helpful information for them. In this part, we compare and discuss several existing microbial disease databases and their extraction methods. Table 3 shows the difference between three databases in microbe and disease data.

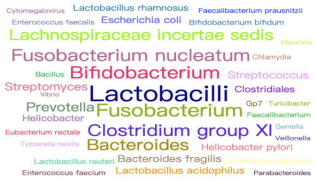
HMDAD (http://www.cuilab.cn/hmdad)[11]: This is the first database of microbe and disease association, the data were collected by manual work, the scope of microbes, diseases and even literature are limited.

Disbiome (https://disbiome.ugent.be)[27]: Didbiome provides a database of the association between the health situation of the host and the composition of its microbiota, it collects microbe-disease associations by text mining from peer-reviewed publications.

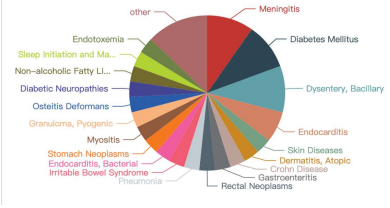
MicroPhenoDB (http://www.liwzlab.cn/microphenodb)[16]: This database uses manual review and calculation methods to systematically integrate the associ-



(a) The pie chart result of disease "Colonic Neoplasms"



(b) The word cloud chart result of disease "Colonic Neoplasms"



(c) The pie chart result of microbe "Bacillus cereus"

(d) The word cloud chart result of microbe "Bacillus cereus"

Figure 7: Example searching result of disease "Colonic Neoplasms" and microbe "Bacillus cereus" in MDIDB.

ated data of pathogenic microorganisms, microbial core genes, and human disease phenotypes. The scoring model is optimized by assigning different weights to different research shreds of evidence to quantify the correlation between microorganisms and human diseases.

Though MicroPhenoDB is rich in data, it takes a lot of time and effort to evaluate and audit the data manually. Moreover, most of the articles may not locate specific sentences described the relations.

MDIDB includes a vast amount of text-mined information from a comprehensive collection of related literature. It also provides a structured way to present the classified relationship between microbial diseases and specific sentences in specific literature.

Our system only contains 1,065 microbial entities, which is due to the lack of specification in the microbial dictionary, and many abbreviated microorganisms in the article, such as B. fragilis, can not be recognized in the NER stage. For the current version, we only consider the microbe disease relationship at sentence level. In the future, we will add relation extraction across sentences.

Table 3: Database contents of MDIDB compared with other databases

	microbe	disease	record	publication	method
HMDAD	292	39	673	61	traditional method
Disbiome	1622	372	10934	1194	traditional method
MicroPhenoDB	**1781**	542	5677	1150	traditional method+manual work
MDIDB	1065	**1198**	**44900**	**8458**	**NLP+deep learning+transfer learning**

Conclusion

Interactions of microbes and diseases are of great importance in the biomedical domain. A lot of valuable information is buried in the large-scale biomedical literature, which has not yet been effectively explored. In this work, we applied text mining to automatically detect the interaction between microbes and diseases from literature via a transfer learning framework. We manually annotated a gold-standard corpus. Then we utilized a state-of-art automated biomedical relation extraction model and fine-tune it on the GSC. The introduction of an automatically generated corpus SSC greatly enlarges the number of training samples and led to a satisfactory performance of 73.85% F_1-score. We conducted five-fold experiments to verify the effectiveness of our transfer learning method, and it provides approximately 10% reduction in error of F_1 score. A total number of 44,900 interactions were extracted from over 20,000 articles. We randomly sampled 1,000 results to analyze the accuracy of the predicted data, and 731 of 1,000 were confirmed correct manually.

Extraction results were utilized to construct a microbe-disease interaction database with a web interface, which is freely available at http://dbmdi.com/index/. Our framework allows large-scale analysis on microbe-disease interactions with evidence of complex sentences.

Acknowledgements
Not applicable.

Funding
This work is jointly funded by the National Science Foundation of China (U1811462), the National Key RD project by Ministry of Science and Technology of China (2018YFB1003203), and the open fund from the State Key Laboratory of High Performance Computing (No. 201901-11). The funder CW took part in the formulation and development of methodology, and provided financial support for this study.

Abbreviations
MDI: Microbe-disease interactions; GSC: Gold standard corpora; SSC: Silver standard corpora ;DNN: Deep neural network; POS: Part of speech

Availability of data and materials
The website is available at http://dbmdi.com/index/.The datasets used and analysed during the current study available from the website.

Ethics approval and consent to participate
No ethics approval and consent were required for the study.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
No ethics approval and consent were required for the study.

Authors' contributions
CW and XX developed the framework and drafted the manuscript; they developed the codes, prepared the datasets for testing, drafted the discussion and revised the whole manuscript together with CY and JC. XX, JY and YQ built the MDIDB website. All authors have read and approved the manuscript.

Author details
[1]State Key Laboratory of High-Performance Computing, National University of Defense Technology, 410073, Changsha, China. [2]College of Computer, National University of Defense Technology, 410073, Changsha, China. [3]Department of General Surgery, Xiangya Hospital, Central South University, 410008, Changsha, China.

References

1. McFarland, L.V.: Beneficial microbes: health or hazard? European journal of gastroenterology & hepatology **12**(10), 1069–1071 (2000)
2. Minakshi, R., Padhan, K., Rehman, S., Hassan, M.I., Ahmad, F.: The sars coronavirus 3a protein binds calcium in its cytoplasmic domain. Virus research **191**, 180–183 (2014)
3. Moni, M.A., Liò, P.: Network-based analysis of comorbidities risk during an infection: Sars and hiv case studies. BMC bioinformatics **15**(1), 333 (2014)

4. Authority, E.F.S., for Disease Prevention, E.C., Control, for Avian influenza, E.U.R.L., Brown, I., Mulatti, P., Smietanka, K., Staubach, C., Willeberg, P., Adlhoch, C., Candiani, D., *et al.*: Avian influenza overview october 2016–august 2017. EFSA Journal **15**(10), 05018 (2017)
5. Bäckhed, F., Fraser, C.M., Ringel, Y., Sanders, M.E., Sartor, R.B., Sherman, P.M., Versalovic, J., Young, V., Finlay, B.B.: Defining a healthy human gut microbiome: current concepts, future directions, and clinical applications. Cell host & microbe **12**(5), 611–622 (2012)
6. Norman, J.M., Handley, S.A., Baldridge, M.T., Droit, L., Liu, C.Y., Keller, B.C., Kambal, A., Monaco, C.L., Zhao, G., Fleshner, P., *et al.*: Disease-specific alterations in the enteric virome in inflammatory bowel disease. Cell **160**(3), 447–460 (2015)
7. Clapp, M., Aurora, N., Herrera, L., Bhatia, M., Wilen, E., Wakefield, S.: Gut microbiota's effect on mental health: the gut-brain axis. Clinics and practice **7**(4), 131–136 (2017)
8. Tran, N., Zhebrak, M., Yacoub, C., Pelletier, J., Hawley, D.: The gut-brain relationship: Investigating the effect of multispecies probiotics on anxiety in a randomized placebo-controlled trial of healthy young adults. Journal of affective disorders **252**, 271–277 (2019)
9. Pasini, E., Aquilani, R., Testa, C., Baiardi, P., Angioletti, S., Boschi, F., Verri, M., Dioguardi, F.: Pathogenic gut flora in patients with chronic heart failure. JACC: Heart Failure **4**(3), 220–227 (2016)
10. Chen, J., Wright, K., Davis, J.M., Jeraldo, P., Marietta, E.V., Murray, J., Nelson, H., Matteson, E.L., Taneja, V.: An expansion of rare lineage intestinal microbes characterizes rheumatoid arthritis. Genome medicine **8**(1), 1–14 (2016)
11. Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., Geng, B., Yang, J., Kong, W., Zhou, X., Cui, Q.: An analysis of human microbe–disease associations. Briefings in bioinformatics **18**(1), 85–97 (2017)
12. Chen, X., Huang, Y.-A., You, Z.-H., Yan, G.-Y., Wang, X.-S.: A novel approach based on katz measure to predict associations of human microbiota with non-infectious diseases. Bioinformatics **33**(5), 733–739 (2017)
13. Bao, W., Jiang, Z., Huang, D.-S.: Novel human microbe-disease association prediction using network consistency projection. BMC bioinformatics **18**(16), 543 (2017)
14. Qu, J., Zhao, Y., Yin, J.: Identification and analysis of human microbe-disease associations by matrix decomposition and label propagation. Frontiers in Microbiology **10**, 291 (2019)
15. Peng, L., Shen, L., Liao, L., Liu, G., Zhou, L.: Rnmfmda: A microbe-disease association identification method based on reliable negative sample selection and logistic matrix factorization with neighborhood regularization. Frontiers in microbiology **11** (2020)
16. Yao, G., Zhang, W., Yang, M., Yang, H., Wang, J., Zhang, H., Wei, L., Xie, Z., Li, W.: Microphenodb associates metagenomic data with pathogenic microbes, microbial core genes, and human disease phenotypes. Genomics, Proteomics & Bioinformatics (2021)
17. Park, Y., Lee, J., Moon, H., Choi, Y.S., Rho, M.: Discovering microbe-disease associations from the literature using a hierarchical long short-term memory network and an ensemble parser model. Scientific reports **11**(1), 1–12 (2021)
18. Leaman, R., Islamaj Doğan, R., Lu, Z.: Dnorm: disease name normalization with pairwise learning to rank. Bioinformatics **29**(22), 2909–2917 (2013)
19. Tanabe, L., Wilbur, W.J.: Tagging gene and protein names in biomedical text. Bioinformatics **18**(8), 1124–1132 (2002)
20. Yeh, A., Morgan, A., Colosimo, M., Hirschman, L.: Biocreative task 1a: gene mention finding evaluation. BMC bioinformatics **6**(S1), 2 (2005)
21. Gerner, M., Nenadic, G., Bergman, C.M.: Linnaeus: a species name identification system for biomedical literature. BMC bioinformatics **11**(1), 85 (2010)
22. Dang, T.H., Le, H.-Q., Nguyen, T.M., Vu, S.T.: D3ner: biomedical named entity recognition using crf-bilstm improved with fine-tuned embeddings of various linguistic information. Bioinformatics **34**(20), 3539–3546 (2018)
23. Zhou, D., Zhong, D., He, Y.: Biomedical relation extraction: from binary to complex. Computational and mathematical methods in medicine **2014** (2014)
24. Semwal, T., Yenigalla, P., Mathur, G., Nair, S.B.: A practitioners' guide to transfer learning for text classification using convolutional neural networks. In: Proceedings of the 2018 SIAM International Conference on Data Mining, pp. 513–521 (2018). SIAM
25. Giorgi, J.M., Bader, G.D.: Transfer learning for biomedical named entity recognition with neural networks. Bioinformatics **34**(23), 4087–4094 (2018)
26. Noronha, A., Modamio, J., Jarosz, Y., Guerard, E., Sompairac, N., Preciat, G., Daníelsdóttir, A.D., Krecke, M., Merten, D., Haraldsdóttir, H.S., *et al.*: The virtual metabolic human database: integrating human and gut microbiome metabolism with nutrition and disease. Nucleic acids research **47**(D1), 614–624 (2019)
27. Janssens, Y., Nielandt, J., Bronselaer, A., Debunne, N., Verbeke, F., Wynendaele, E., Van Immerseel, F., Vandewynckel, Y.-P., De Tré, G., De Spiegeleer, B.: Disbiome database: linking the microbiome to disease. BMC microbiology **18**(1), 1–6 (2018)
28. Song, M., Kim, W.C., Lee, D., Heo, G.E., Kang, K.Y.: Pkde4j: Entity and relation extraction for public knowledge discovery. Journal of biomedical informatics **57**, 320–332 (2015)
29. Hong, L., Lin, J., Li, S., Wan, F., Yang, H., Jiang, T., Zhao, D., Zeng, J.: A novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories. Nature Machine Intelligence, 1–9 (2020)
30. Ramarao, N., Tran, S.-L., Marin, M., Vidic, J.: Advanced methods for detection of bacillus cereus and its pathogenic factors. Sensors **20**(9), 2667 (2020)
31. Ehling-Schulz, M., Lereclus, D., Koehler, T.M.: The bacillus cereus group: Bacillus species with pathogenic potential. Gram-Positive Pathogens, 875–902 (2019)
32. Ehling-Schulz, M., Frenzel, E., Gohar, M.: Food–bacteria interplay: pathometabolism of emetic bacillus cereus. Frontiers in microbiology **6**, 704 (2015)
33. Stevens, M.P., Elam, K., Bearman, G.: Meningitis due to bacillus cereus: a case report and review of the

literature. Canadian Journal of Infectious Diseases and Medical Microbiology **23**(1), 16–19 (2012)
34. Orrett, F.: Fatal bacillus cereus bacteremia in a patient with diabetes. Journal of the National Medical Association **92**(4), 206 (2000)
35. Mohammadi, G., Adorian, T.J., Rafiee, G.: Beneficial effects of bacillus subtilis on water quality, growth, immune responses, endotoxemia and protection against lipopolysaccharide-induced damages in oreochromis niloticus under biofloc technology system. Aquaculture Nutrition **26**(5), 1476–1492 (2020)
36. Hendler, R., Zhang, Y.: Probiotics in the treatment of colorectal cancer. Medicines **5**(3), 101 (2018)
37. Keku, T.O., McCoy, A.N., Azcarate-Peril, A.M.: Fusobacterium spp. and colorectal cancer: cause or consequence? Trends in microbiology **21**(10), 506–508 (2013)
38. Parisa, A., Roya, G., Mahdi, R., Shabnam, R., Maryam, E., Malihe, T.: Anti-cancer effects of bifidobacterium species in colon cancer cells and a mouse model of carcinogenesis. PloS one **15**(5), 0232930 (2020)
39. Rubinstein, M.R., Baik, J.E., Lagana, S.M., Han, R.P., Raab, W.J., Sahoo, D., Dalerba, P., Wang, T.C., Han, Y.W.: Fusobacterium nucleatum promotes colorectal cancer by inducing wnt/β-catenin modulator annexin a1. EMBO reports **20**(4), 47638 (2019)
40. Abed, J., Maalouf, N., Manson, A.L., Earl, A.M., Parhi, L., Emgård, J.E., Klutstein, M., Tayeb, S., Almogy, G., Atlan, K.A., et al.: Colon cancer-associated fusobacterium nucleatum may originate from the oral cavity and reach colon tumors via the circulatory system. Frontiers in cellular and infection microbiology **10** (2020)
41. Guarner, F., Malagelada, J.-R.: Gut flora in health and disease. The Lancet **361**(9356), 512–519 (2003)
42. Moore, W., Moore, L.H.: Intestinal floras of populations that have a high risk of colon cancer. Applied and environmental microbiology **61**(9), 3202–3207 (1995)
43. Vacca, M., Celano, G., Calabrese, F.M., Portincasa, P., Gobbetti, M., De Angelis, M.: The controversial role of human gut lachnospiraceae. Microorganisms **8**(4), 573 (2020)
44. Cueva, C., Silva, M., Pinillos, I., Bartolomé, B., Moreno-Arribas, M.: Interplay between dietary polyphenols and oral and gut microbiota in the development of colorectal cancer. Nutrients **12**(3), 625 (2020)
45. Bolourian, A., Mojtahedi, Z.: Streptomyces, shared microbiome member of soil and gut, as 'old friends' against colon cancer. FEMS microbiology ecology **94**(8), 120 (2018)
46. Boleij, A., Schaeps, R.M., Tjalsma, H.: Association between streptococcus bovis and colon cancer. Journal of clinical microbiology **47**(2), 516–516 (2009)

Figures

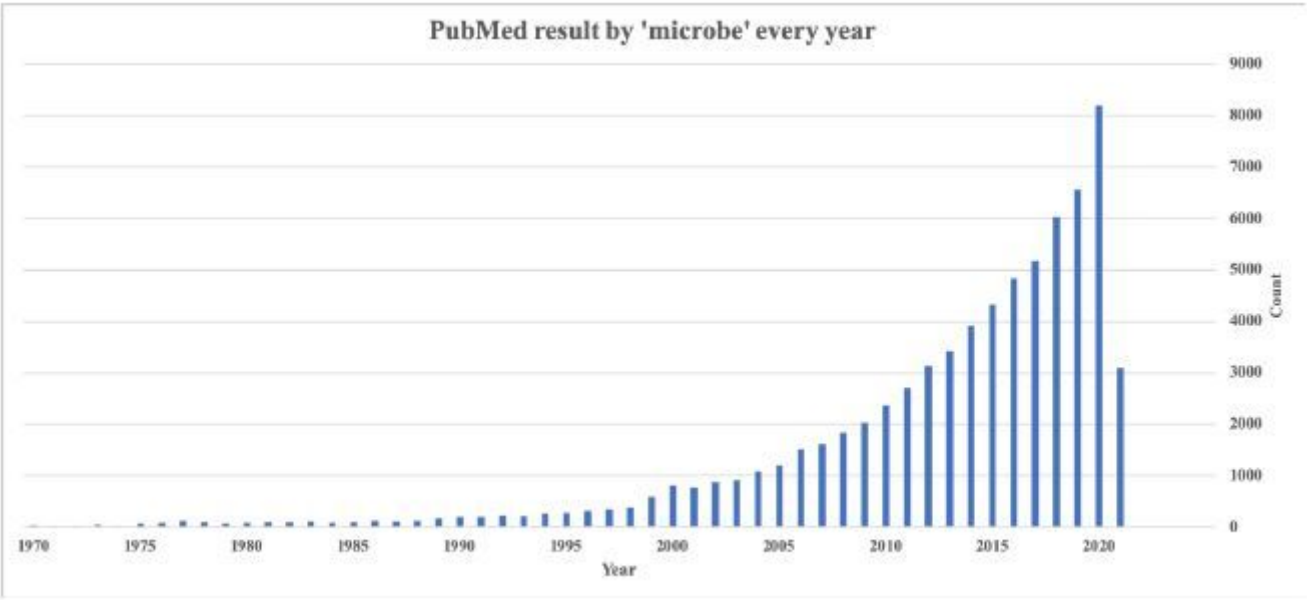


Figure 1

Number of PubMed articles returned by the `microbe' query.

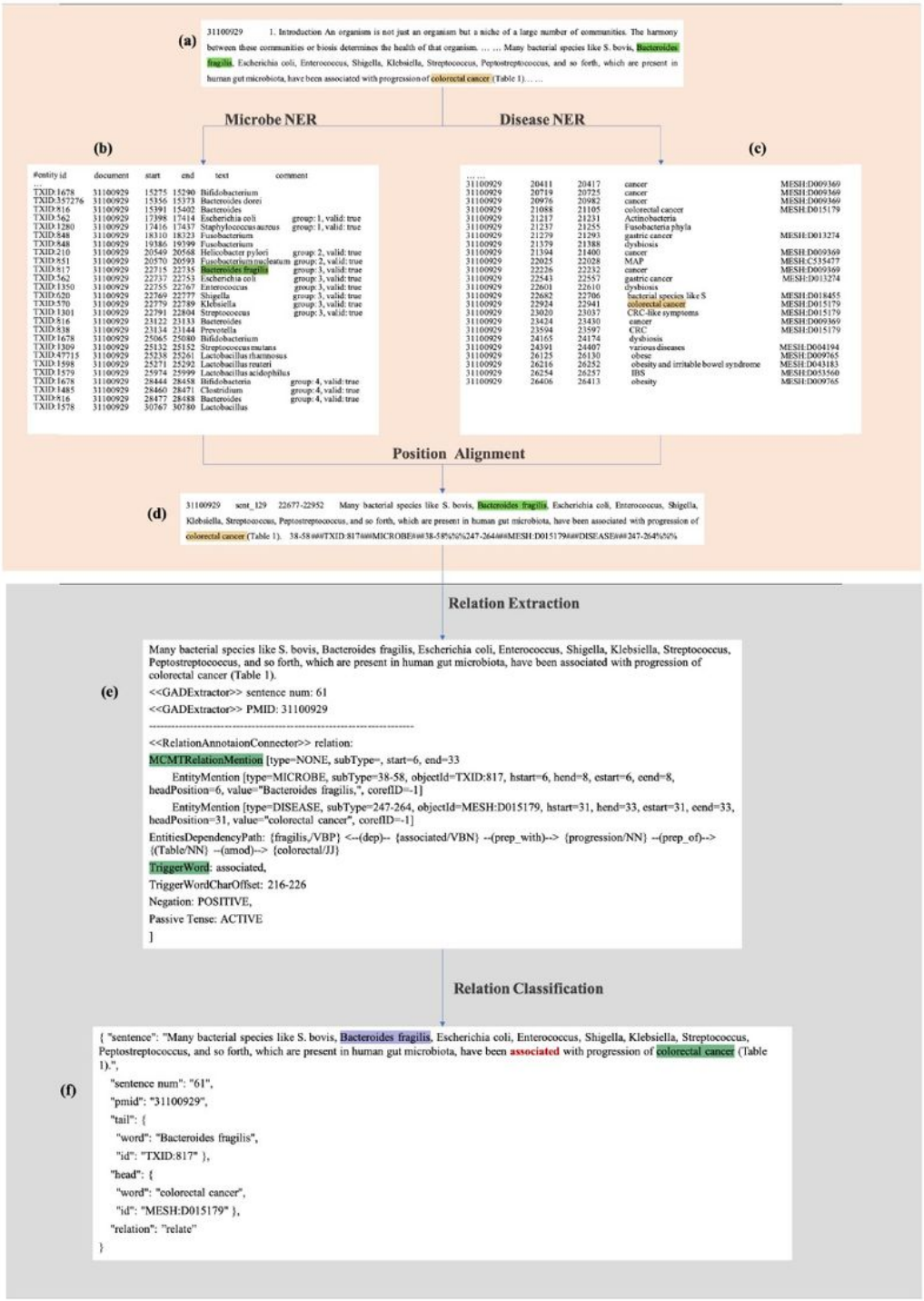


Figure 2

The workflow of data preparation. (a)The initial full-text data (b)The result of applying LINNAEUS to recognize microbe entities (c)The result of applying DNORM to recognize disease entities (d)The result of splitting texts into sentences and aligning the position of disease and microbe (e)The result of relation extraction using PKDE4J (f)The result of classification instance by the result of PKDE4J

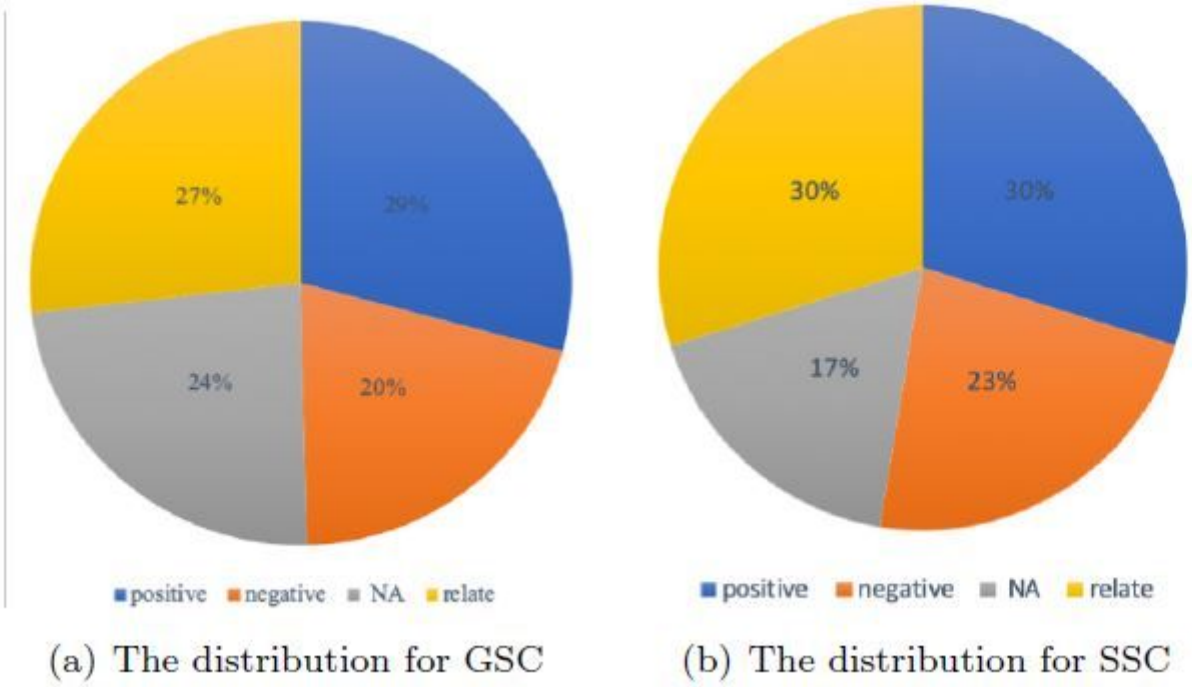


Figure 3

The distribution of relation types for GSC and SSC

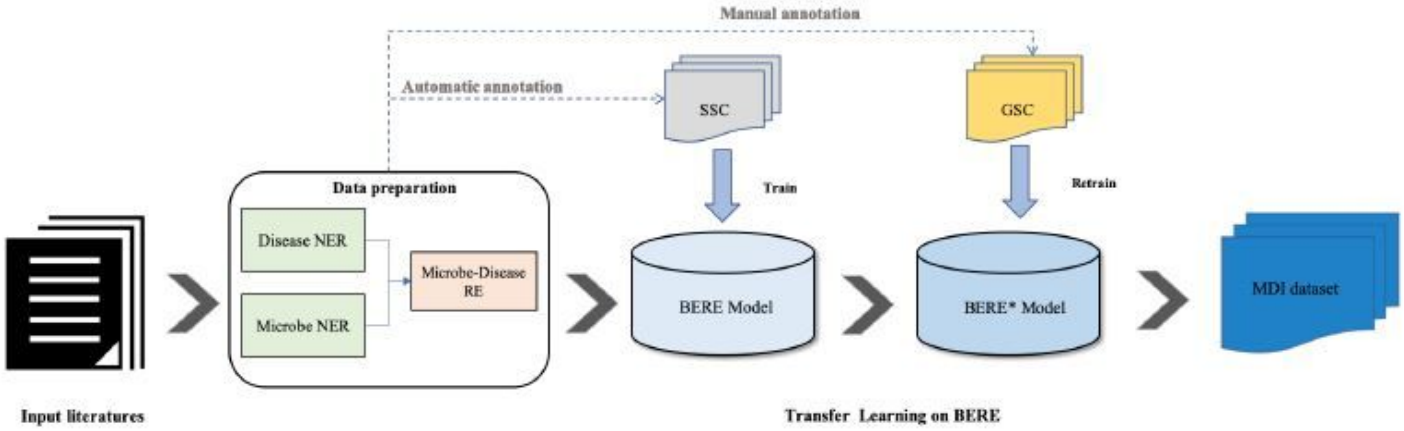


Figure 4

The procedures of mining information from literature

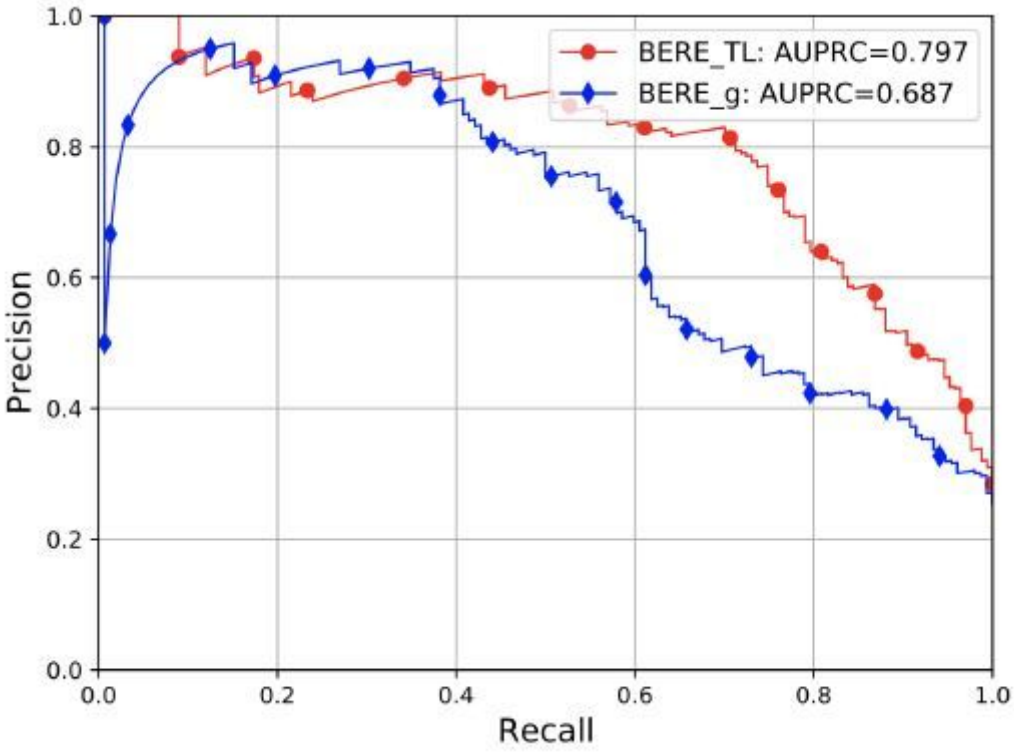
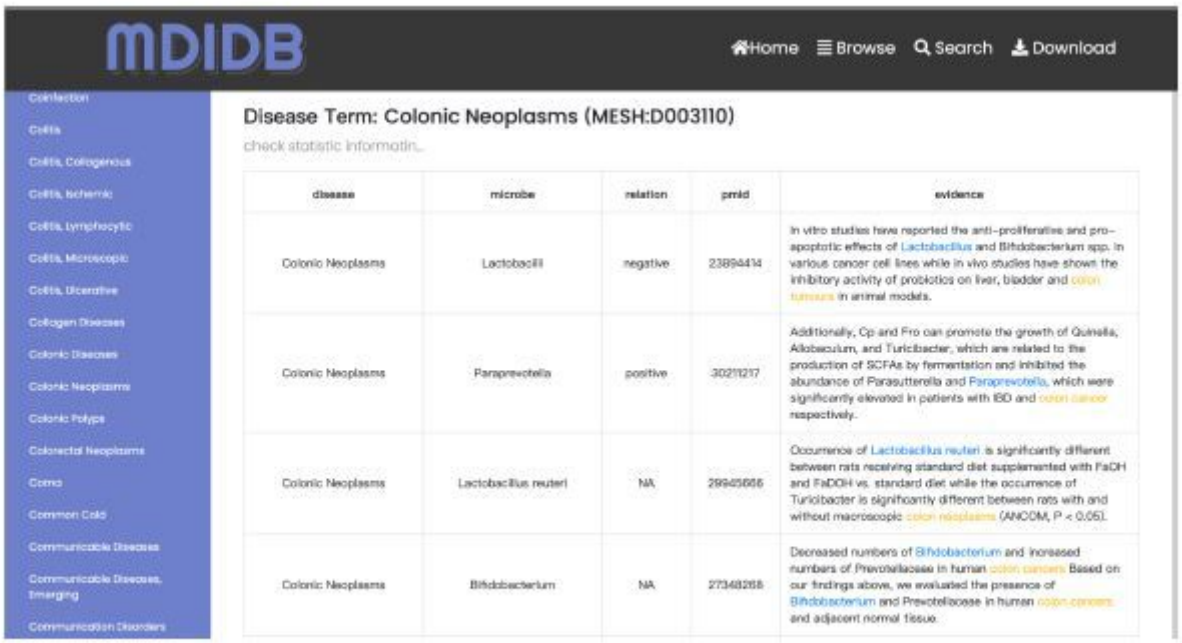
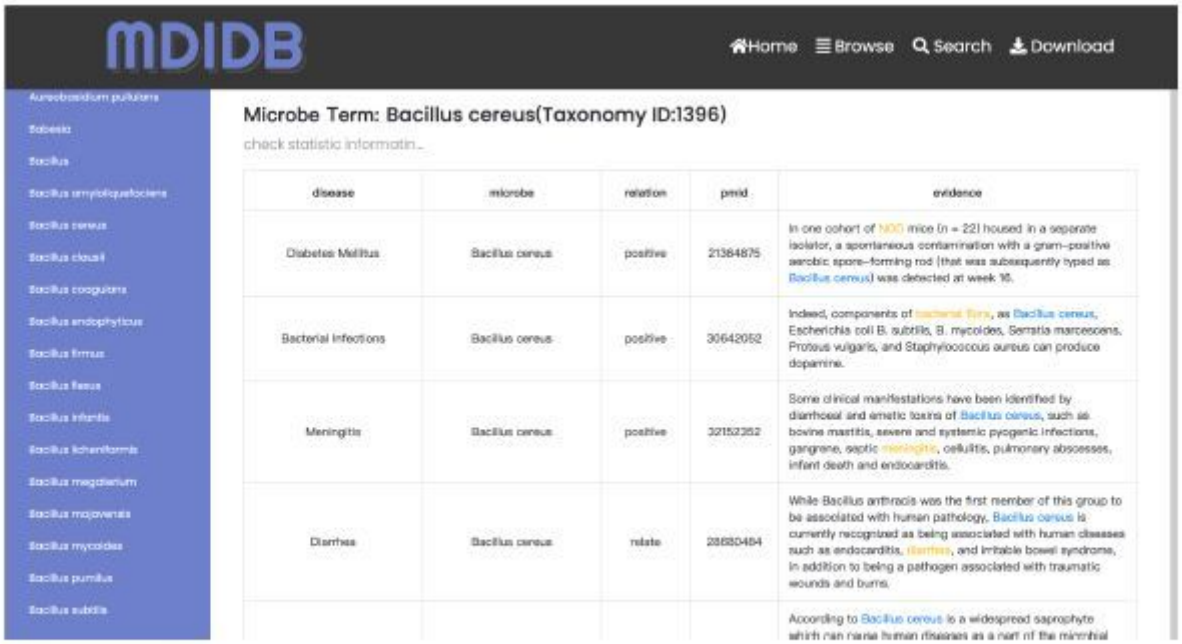


Figure 5

Comparisons of the Precision{Recall curves between BERE with or without transfer learning. The AUPRC and F1-score for each method are on the top right contains.



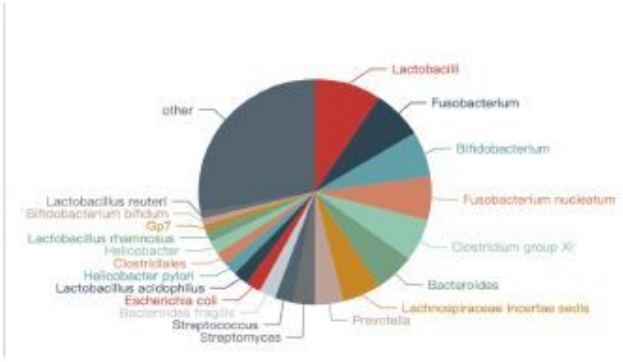
(a) Searching result of disease "Colonic Neoplasms"



(b) Searching result of microbe "Bacillus cereus"

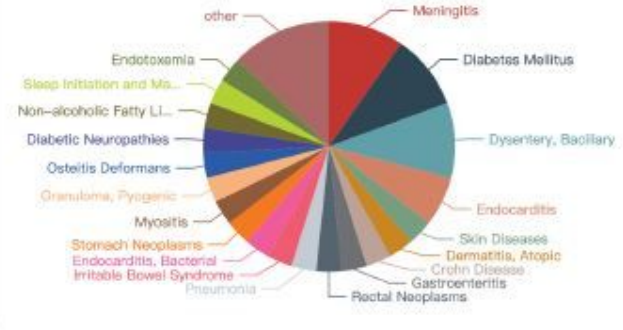
Figure 6

Query results in MDIDB.



(a) The pie chart result of disease "Colonic Neoplasms"

(b) The word cloud chart result of disease "Colonic Neoplasms"



(c) The pie chart result of microbe "Bacillus cereus"

(d) The word cloud chart result of microbe "Bacillus cereus"

Figure 7

Example searching result of disease "Colonic Neoplasms" and microbe "Bacillus cereus" in MDIDB.