PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Scaling up data curation using deep learning: An application to literature triage in genomic variation resources

Kyubum Lee[1], Maria Livia Famiglietti[2☯], Aoife McMahon[3☯], Chih-Hsuan Wei[1], Jacqueline Ann Langdon MacArthur[3], Sylvain Poux[2], Lionel Breuza[2], Alan Bridge[2], Fiona Cunningham[3], Ioannis Xenarios[4,5], Zhiyong Lu[1]*

1 National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, Maryland, United States of America, **2** Swiss-Prot Group, SIB Swiss Institute of Bioinformatics, Geneva, Switzerland, **3** European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom, **4** Center for Integrative Genomics, University of Lausanne, Lausanne Switzerland, **5** Department of Chemistry and Biochemistry, University of Geneva, Geneva, Switzerland

☯ These authors contributed equally to this work.
* zhiyong.lu@nih.gov

OPEN ACCESS

Citation: Lee K, Famiglietti ML, McMahon A, Wei C-H, MacArthur JAL, Poux S, et al. (2018) Scaling up data curation using deep learning: An application to literature triage in genomic variation resources. PLoS Comput Biol 14(8): e1006390. https://doi.org/10.1371/journal.pcbi.1006390

Editor: Rong Xu, Case Western Reserve University, UNITED STATES

Received: April 5, 2018

Accepted: July 24, 2018

Published: August 13, 2018

Data Availability Statement: We used publicly available data from PubMed, the GWAS Catalog and UniProtKB/SwissProt. We included the download URLs as Supporting Information.

Funding: This research was supported by the National Institutes of Health Intramural Research Program, National Library of Medicine. Swiss-Prot group activities are supported by the Swiss Federal Government through the State Secretariat for Education, Research and Innovation (SERI) and by the National Institutes of Health (NIH) (UniProt

Abstract

Manually curating biomedical knowledge from publications is necessary to build a knowledge based service that provides highly precise and organized information to users. The process of retrieving relevant publications for curation, which is also known as document triage, is usually carried out by querying and reading articles in PubMed. However, this query-based method often obtains unsatisfactory precision and recall on the retrieved results, and it is difficult to manually generate optimal queries. To address this, we propose a machine-learning assisted triage method. We collect previously curated publications from two databases UniProtKB/Swiss-Prot and the NHGRI-EBI GWAS Catalog, and used them as a gold-standard dataset for training deep learning models based on convolutional neural networks. We then use the trained models to classify and rank new publications for curation. For evaluation, we apply our method to the real-world manual curation process of UniProtKB/Swiss-Prot and the GWAS Catalog. We demonstrate that our machine-assisted triage method outperforms the current query-based triage methods, improves efficiency, and enriches curated content. Our method achieves a precision 1.81 and 2.99 times higher than that obtained by the current query-based triage methods of UniProtKB/Swiss-Prot and the GWAS Catalog, respectively, without compromising recall. In fact, our method retrieves many additional relevant publications that the query-based method of UniProtKB/Swiss-Prot could not find. As these results show, our machine learning-based method can make the triage process more efficient and is being implemented in production so that human curators can focus on more challenging tasks to improve the quality of knowledge bases.

5U41HG007822-02). Research reported in this publication was also supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number U41HG007823. This research was also supported by the European Molecular Biology Laboratory. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

As the volume of literature on genomic variants continues to grow at an increasing rate, it is becoming more difficult for a curator of a variant knowledge base to keep up with and curate all the published papers. Here, we suggest a deep learning-based literature triage method for genomic variation resources. Our method achieves state-of-the-art performance on the triage task. Moreover, our model does not require any laborious preprocessing or feature engineering steps, which are required for traditional machine learning triage methods. We applied our method to the literature triage process of UniProtKB/Swiss-Prot and the NHGRI-EBI GWAS Catalog for genomic variation by collaborating with the database curators. Both the manual curation teams confirmed that our method achieved higher precision than their previous query-based triage methods without compromising recall. Both results show that our method is more efficient and can replace the traditional query-based triage methods of manually curated databases. Our method can give human curators more time to focus on more challenging tasks such as actual curation as well as the discovery of novel papers/experimental techniques to consider for inclusion.

Introduction

The question of how genetic variation in a population influences phenotypic variation is of major importance in biology. Naturally occurring genetic variants, both rare and common, can provide insight into disease mechanism and protein function. This understanding, coupled with the recent explosion in next-generation sequencing, has meant a dramatic increase in publications on the subject. This also means that it is now impossible for individual researchers to collect and collate all the variant information that may be relevant to them. To assist researchers, variant information in publications is selected, summarized, organized, and stored in a searchable form in knowledge bases such as UniProtKB/Swiss-Prot [1, 2] and the NHGRI-EBI GWAS Catalog [3] for easier access and greater usability. However, such knowledge bases require domain experts to collect and manually curate high quality information from the literature [4], a highly time consuming and therefore costly process. In addition, these knowledge bases have their own focus and collect information from publications according to their own specific guidelines.

As shown in the study of Baumgartner et al., the manual curation of information for genomic databases is, for the most part, not scalable due to the fast-growing number of publications [5]. To overcome this scalability issue, automated or semi-automated methods can be used [6]. For this purpose, text mining and machine learning tools that support information extraction and annotation have been developed [7–14]. Poux et al. [15] demonstrated that manual curation can be more efficient and scalable with the proper text-mining services and techniques.

As a typical first step in the manual curation process, the document triage process involves identifying publications of interest [16, 17]. On average, three thousand biomedical papers are indexed in PubMed every day. (In 2016, more than 1.1 million publications were indexed in PubMed). Triage is then required to select a subset of relevant publications for curation. As Hirschman et al. [16] explained, triage is usually carried out using pre-defined queries in the PubMed database. The queries are generated using general terms related to topics such as "GWAS" and "Drug screening" or a list of entities (e.g., onco-gene list [18]). Publication date or publication type (e.g., review, book) are also included for a query as needed. However, a triage process using pre-defined queries has several major limitations. Using general topic terms for a

query may be futile because publications on a certain topic may not contain the specific terms. For example, if a user wishes to find publications on cancer-related genes, the query should contain specific terms such as "HER2," "Carcinoma," "Tumor," and so on, rather than "cancer-related gene" because many publications on cancer-related genes do not contain the general term "cancer-related gene." Furthermore, the topic words in a publication may provide only background information and may not be on the main topic of the publication. Also, longstanding databases have their own complex and detailed guidelines for manual curation and it can be difficult to create a PubMed query that fulfills their conditions [19]. For these reasons, query-based triage is limited in retrieving highly precise and complete set of relevant publications for further use, not to mention the process itself is also very labor intensive and time consuming.

To overcome the limitations of the query-based triage and manual curation processes, machine learning-assisted curation research studies have been conducted. Poux et al. [15] used PubTator, a web-based curation support system that assists users in annotating publications in PubMed. Curators can read publications that are highlighted and pre-annotated by automated named-entity recognition tools. Curators can also easily upload and generate their curation collection and save their results with a simple mouse click. In their study, Poux et al. selected only thirteen journals from which to collect protein function information, and ranked the publications of the journals by the number of proteins mentioned in the text. However, other than the thirteen selected journals, there are many journals that have published papers on protein function. Also, prioritizing the publications by the number of proteins may not be the best method because some papers include valuable information on a small number of proteins.

Almeida et al. [20] built a method called mycoSORT using support vector machine (SVM), Naïve Bayes and Logistic Model Trees on the triage task for the mycoCLAP database [21]; however, this required several text preprocessing steps and an extensive feature extraction process which are data/domain dependent. Because of these dependencies, their method is not directly applicable to other types of databases. Their feature extraction process is time consuming, labor intensive, and requires domain knowledge from human experts [6].

In recent years, newly proposed deep learning-based text mining methods have started outperforming traditional machine learning-based methods in various tasks [22–28]. In addition, these deep neural network models do not require intensive feature engineering by domain experts; hence, they can also be easily generalized to other tasks with datasets in different domains. In this paper, we propose to employ convolutional neural network (CNN), a class of deep, feed-forward artificial neural networks, for the identification of publications relevant for variant curation. By comparing the results of our method with those of mycoSORT, the method proposed by Almeida et al. [20], we demonstrate how our deep learning-based classifier performs better than traditional machine learning classifiers, even without feature engineering. For assessing the utility of our approach, we applied our method to two external knowledge bases (UniProtKB/Swiss-Prot and the GWAS Catalog) in real-world circumstances. We compare the performance of our proposed method with that of each of the knowledge base-specific query-based methods to demonstrate that our method can greatly improve the efficiency of the document triage step in these two databases. While the usage of CNNs to classify text documents is not new, the application of deep learning to speed up the real-world triage process for biomedical literature is, to the best of our knowledge.

Results

In this research, our goal is to improve the triage process by predicting the most suitable publications for each knowledge base in their manual curation. We aim to provide both binary and ranked results with scores for each selected publication.

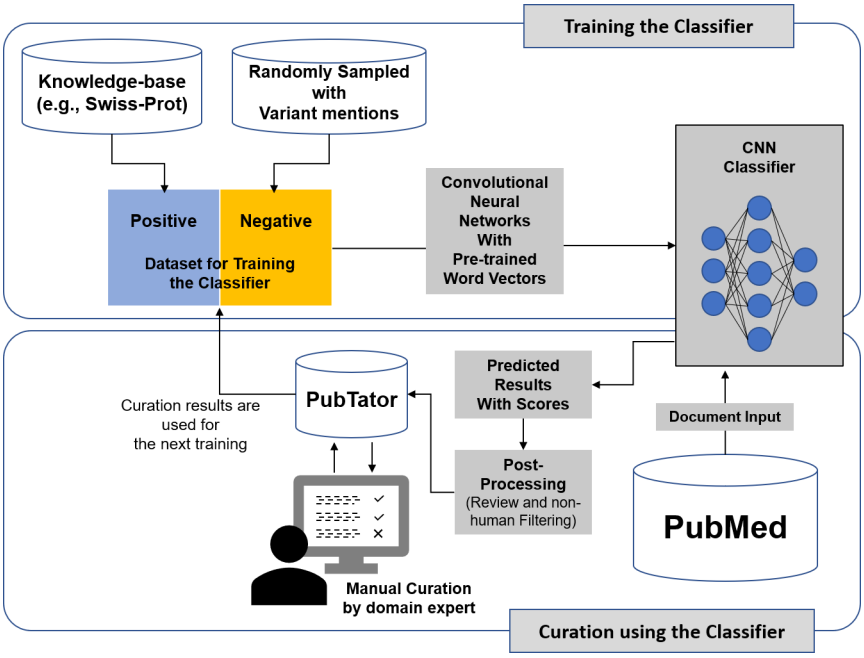

Fig 1. Literature triage using our deep learning framework.

https://doi.org/10.1371/journal.pcbi.1006390.g001

Fig 1 shows the overview of our proposed framework. In a nutshell, using the previously curated publications from each knowledge base as positive examples and other variant-containing publications as negatives, we first train our machine learning classifiers. Then, we classify and rank new publications in PubMed using the trained classifier. Finally, we import classified results into PubTator, and provide the results to curators for manual verification.

For the evaluation of our method, we collected publications from UniProtKB/Swiss-Prot [1, 2] and the GWAS Catalog [3], both of which are widely used variant information knowledge bases containing several thousand manually curated publications. For method comparison, we also collected a previous document triage dataset called mycoSet [20] where other machine learning approaches were tested.

Document classification results

We evaluated our deep learning-based triage method on three different datasets. Table 1 shows the performance of our method on the publications in UniProtKB/Swiss-Prot and the

Table 1. Classification performance on UniProtKB, the GWAS Catalog and mycoSet. (CNN: Convolutional Neural Networks, SVM: Support Vector Machine, LMT: Logistic Model Trees).

Dataset	Methods	Precision	Recall	F1
UniProtKB/Swiss-Prot	Our Method (CNN)	0.913	0.934	**0.923**
	LinearSVC (SVM)	0.896	0.920	0.908
The GWAS Catalog	Our Method (CNN)	0.973	0.991	**0.982**
	LinearSVC (SVM)	0.965	0.980	0.972
mycoSet*	Our Method (CNN)	0.602	0.667	**0.633**
	LinearSVC (SVM)	0.566	0.627	0.595
	mycoSORT (LMT)	0.552	0.6	0.575

* In this table, the positive vs. negative ratio of mycoSet is 1:9, and that of the other datasets is 1:1.

https://doi.org/10.1371/journal.pcbi.1006390.t001

GWAS Catalog published before 2017 and mycoSet. Overall, our method achieved high performance on this triage task. We hypothesized that our method would achieve higher performance on the GWAS Catalog dataset than the other two datasets because most of the GWAS-related publications contain terms such as "GWAS" or "Genome-wide association".

We compared our classification results with those of the LinearSVC classifier which is a non-deep learning-based method and has achieved the best results in a recent text classification task [29]. LinearSVC is a classifier based on Support Vector Machine with a linear kernel. We used the implementation of LinearSVC with Lasso regression in Scikit-learn [30] for Python.

We also compared our classification results with those of mycoSORT [20]. Our method achieved higher precision, recall, and F1 scores than mycoSORT [20], which uses traditional machine learning techniques such as SVM, Naïve Bayes and Logistic Model Trees, as shown in Table 1. It is important to note that our model did not use any preprocessing, information tagging, or feature engineering steps, all of which were used in mycoSORT.

Notice that the ratio of positive vs. negative instances of mycoSet is set to 1:4 for the training and validation sets, and 1:9 for the test set in order to be consistent with Almeida et al.'s mycoSORT [20] evaluation settings. The other datasets have 1:1 ratios for their training, validation and test sets. Note that in practice, those ratios are quite different in the triage process (see details in the section on imbalanced dataset below).

Word distribution in the datasets

To better understand the classification results, we obtained lists of the enriched words from the positively classified publications. We counted the number of times each word was mentioned in the positive and negative publications and obtained the top 100 most statistically significant words from each dataset using chi square test. Table 2 shows the 26 most significant terms in each dataset after manually combining the different forms of the same word (e.g. plural and past tense). An interesting observation is that documents in the UniProtKB/Swiss-Prot dataset often contain the word "mutation" while documents in the GWAS Catalog contain "variants" or "SNP(s)" at the top of its list. The words that overlap in the queries of each databases' query-based triage method are highlighted.

Table 2. Lists of the most significant words in the positively classified publications. (The words that are used as queries in the query-based method of each database are highlighted.).

UniProtKB/Swiss-Prot		NHGRI-EBI GWAS Catalog	
mutation(s)	syndrome	wide	variants
gene	exon(s)	genome	meta
cdna	encoding	association(s)	european
human	chromosome	loci (or locus)	identify (-ies, -ied)
sequence	two	snp(s)	susceptibility
missense	region	gwas	near
families	acid	p	ancestry
novel	coding	=	chromosome
amino	domain	10	significance
identified	expressed	genetic	8
family	recessive	study	independent
autosomal	affected	replication (replicated)	conducted
protein	cloning	associated	cohorts

https://doi.org/10.1371/journal.pcbi.1006390.t002

The different purposes of the two different knowledge bases are reflected in the differences in the word lists. Since the UniProtKB/Swiss-Prot positive publications contain protein altering mutations, its word list includes words such as "protein," "amino," "acid," "sequence," "coding," "region," and "missense." Words related to typical GWAS study design and aims (cohort, replication, susceptibility and p-value related tokens such as 'p,' ' = ' and '10') are also found.

The GWAS Catalog results contain terms such as "Genome-wide association study" or "GWAS," which are included in the PubMed query used by the query-based method of the knowledge base. However, the UniProtKB/Swiss-Prot results do not contain the terms that are included in the query which is "functional characterization and functional analysis".

Based on this result, we verify that our method achieves high performance in classifying publications. We also confirmed that the positively classified publications contained the correct signals.

Ranking results of unbalanced datasets in a real-world setting

In a real-world application, we need to address the issue of extremely unbalanced datasets in terms of the ratio between positive and negative documents. Ranked results with scores are helpful in practice for manual curators because they can freely decide the number of documents to curate or discard.

Between January to July 2017, 23 and 225 publications were included by UniProtKB/Swiss-Prot and the GWAS Catalog through manual curation, respectively. Thus, these articles are treated as positives. The other 0.6 million publications are considered as negative. Since these datasets are extremely imbalanced, we evaluated our method in ranking/scoring using receiver operating characteristic (ROC) curves. Fig 2(A) shows the ROC curves of the ranked results by our method. We can see that most of the positive publications are ranked at the very top of the results of the classifiers and achieved high AUROCs such as 0.995 and 0.998 for UniProtKB/Swiss-Prot and the GWAS Catalog, respectively.

Additionally, we plotted Fig 2(B) using only the publications containing variants at the abstract level. We used tmVar [8, 31] to find the variant-containing publications and around

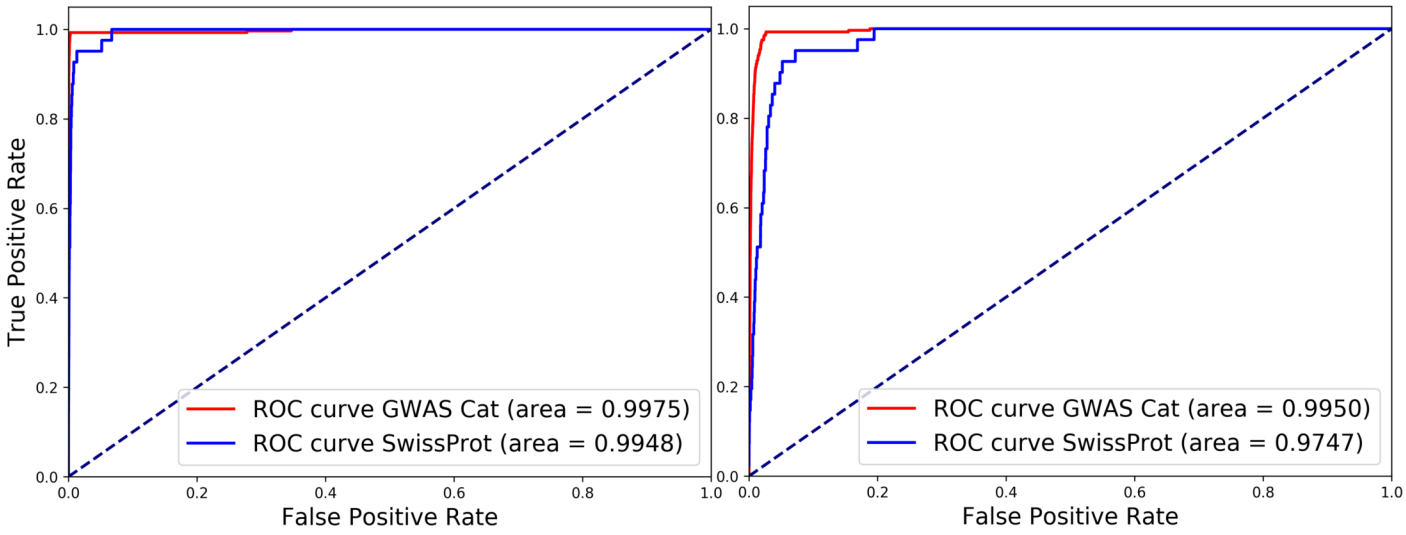

Fig 2. ROC curves of the classification results on the 2017JanJul group of UniProtKB/Swiss-Prot (Blue) and the GWAS Catalog (Red)–(a) Curves in all the publications, (b) Curves in the publications containing mutations at the abstract level.

https://doi.org/10.1371/journal.pcbi.1006390.g002

Table 3. Comparison of the results of our method with those of the query-based method in the UniProtKB/Swiss-Prot and GWAS Catalog triage. Both query-based and CNN-based results were evaluated by the curators, resulting in the total number of curatable publications below.

		UniProtKB/Swiss-Prot	GWAS Catalog
Total number of target publications		4,680 (3 months, variant-containing publications)	64,405 (3 weeks, all publications)
Total number of curatable publications		424	27
Query-based	**Total**	79	304
	Curatable	36 (P: 45.57%, R: 8.49%)	27 (P: 8.88%, R: 100%)
Our CNN-based Method*	**Total**	501	98
	Curatable	413 (P: 82.43%, R: 97.41)	26 (P: 26.53%, R: 96.30%)

* As requested by the curators, UniProtKB results are filtered using tmVar as only articles with explicit variant mentions are within the scope of its data curation.

https://doi.org/10.1371/journal.pcbi.1006390.t003

10,000 of those were found. In this result, the ROCs are slightly lower compared to those shown in Fig 2(A), but still we can see that all the positive publications are ranked in the top of the results even in variant-containing publications.

Utility assessment

For evaluating its utility, our method was applied to the triage process of UniProtKB/Swiss-Prot and the GWAS Catalog, in collaboration with the database curators. The comparison between the query-based method and our method for both databases is provided in Table 3. To evaluate and compare our method's results with those of the query-based method, the curators of both the databases manually curated not only the results of the query-based method but also the publications found only by our method. We consider all the curatable publications found by both methods as positives. The documents not found by either method are considered as negatives, although some might have been deemed as curatable if retrieved. Hence, the recall we reported in this section might be slightly lower than the actual recall. Nonetheless, we still believe our recall is useful for comparing the performance of the two methods.

Machine learning-assisted manual curation framework—PubTator. To manage the curation process more efficiently, we used the PubTator curation system [32]. After users uploading a list of PMIDs for evaluation, PubTator provides users with an organized web-based user curation interface that displays the text of a publication in which biomedical entities are highlighted. Furthermore, users can indicate whether the publications are curatable by clicking a button. The curation results can be saved and downloaded for future use. Manual curation using PubTator has proven to be efficient in a previous study [15].

We generated the prediction results of the databases, uploaded them to PubTator, and created a collection of publications for manual curation. The curators of each database, who are also domain experts, curated the collections and sent their feedback through the PubTator system.

UniProtKB/Swiss-Prot. For UniProtKB/Swiss-Prot, we provided collections of positively predicted articles published in August, September and October 2017. Since the curators of UniProtKB/Swiss-Prot wanted to curate publications containing variant names at the abstract level, we used tmVar [31] to remove publications without variant mentions.

Table 3 shows the comparison of the results of our machine learning-based method and those of the current query-based UniProtKB/Swiss-Prot method. Our machine learning-based method found more publications than their current query-based method. Also, our method achieved a precision of 82.43% while the query-based method obtained a lower precision of 67.74%. Out of all the 425 curatable papers, our method missed only 12 publications (97.17% recall) compared with the query-based method that missed 398 publications (6.35% recall).

Our method found 11 curatable publications that the query-based method could not find. The manual curator of UniProtKB/Swiss-Prot confirmed that our method addresses their need to get a wide set of articles with high precision without returning an overwhelming number of items.

The GWAS Catalog. We predicted publications for the GWAS Catalog for three weeks from January 10, 2018 to January 30, 2018. We compared our results with their query results in Table 3. Since the GWAS Catalog is not interested in non-human research and review papers, we found and removed those publications using PubTator and PubMed meta information.

As shown in Table 3, our method achieved a precision of 26.53% versus a precision of 8.88% obtained by the query-based method. Our method retrieved a smaller number of results which include all the relevant publications except for one (PMID 29313844) found by the traditional query-based method. Thus, our method returns 1/3 of the results returned by the query-based method, which reduces the number of irrelevant publications for manual review by curators. In this case, our method did not find any new publications beyond the query-based results. Also, our method missed a curatable publication that the query-based method found, which is explained in the Discussion section.

Discussion

Our results demonstrate that our machine learning triage method works well on different datasets in different settings. Both the UniProtKB/Swiss-Prot and the GWAS Catalog manual curation teams confirmed that our method achieved higher precision than the previous query-based triage methods without significantly compromising recall. In the UniProtKB/Swiss-Prot result, our method could find many relevant publications that the query-based method could not find. In the GWAS Catalog case, our method significantly improved the efficiency of the curation process by reducing the number of papers that required review. Both results show that our method can replace the traditional query-based triage methods of manually curated databases.

However, even though our method works well and improves the efficiency of the manual curation process, it is difficult for our method to include new types of articles based on new interests. For example, the GWAS Catalog, which has collected publications related to only genome-wide array-based association studies so far, might find it difficult to find whole-exome sequencing publications since our classifier is trained on only the publications related to array-genotyped GWAS and there are no whole-exome sequencing papers in the training set. This is known as the "filter bubble" problem of personalized search, social media and news recommender systems [33]. For example, based on the behavior of a user, personalized news recommendation algorithms suggest news articles on certain topics as results. The user selects some of the results recommended by the algorithm, and the algorithm learns from those results again to recommend similar results. After this process is repeated a few times, the recommendation algorithm narrows down the recommendation results based on the interests of the user, which makes it more difficult for users to view new information other than the results recommended by the algorithm. To solve this problem, the curators need to manually find new topic publications as a separate activity. However, the curators and our team both agreed that our classification and ranking methods would save curators considerable time in standard triage. Thus they have extra time to manually find new subjects, using new queries in PubMed or using their intuition, which can solve the filter bubble problem.

The other limitation of our method comes from the "black box" property of the deep neural network models [34, 35]. In the utility assessment of the GWAS Catalog section, our method

did not retrieve a publication (PMID 29313844) while the query-based method did. The confidence score of the publication obtained by our method was 0.451 which is slightly lower than the default threshold of the classifiers (0.5). We investigated why the publication was scored low even though it contains GWAS relevant terms, such as "p-value" and "GWAS," however we could not clearly understand why this publication was not scored highly enough because deep neural networks are "black boxes." With other traditional machine learning models such as decision tree, k-nearest neighbor and logistic regression, it is possible to see how the classification of each case is made by looking at the decision rules, neighbors, or weight of features, respectively; however, it is not possible in deep neural network models. This example shows the limitation of deep learning methods.

The hardest part of designing our method was that there is no negative gold-standard dataset available for training the classifiers. We could not collect any of the negative data for training / testing our method as we were able to for mycoSet [20]. We constructed negative datasets using randomly collected variant containing publications using tmVar after filtering out the positive publications; however, there are chances that some positive examples are included in the negative dataset. Using tmVar for generating negative dataset may also cause some bias in the classification results, because all the negative publications are tmVar positive. Our method still over-performed compared to the query-based triage methods. If we can collect enough gold-standard negative publications, our method may perform better. Since PubTator is now used as a manual literature triage framework for several knowledge bases such as UniProtKB/Swiss-Prot and the GWAS Catalog, we can easily collect negative publications that curators mark as irrelevant and we expect to use them as negative gold-standard data for training the classifiers in the future.

In this research, we focus on the triage of variant-related knowledge bases; however, our method can be applied to any knowledge base that relies on manual curation and to a triage process if it has a sufficient number of documents for training the classifier.

We also believe it would be interesting to use our method for knowledge bases of biomedical relations such as protein-protein interactions and drug-drug interactions. Although we did not use any preprocessing in our method, named-entity tagging might be helpful in these tasks.

Materials and methods

UniProtKB/Swiss-Prot

UniProtKB/Swiss-Prot [2] is a knowledge base of protein sequence and functional information based on manual curation and is a part of the universal protein knowledge base UniProt [1]. UniProtKB/Swiss-Prot contains rich information on genomic variants that affect protein function [2]. Poux et al. [36] explained that the curation process of UniProtKB/Swiss-Prot is expensive and time consuming. Currently, the triage process of variant information in UniProtKB/Swiss-Prot is performed using manually pre-defined queries in PubMed; however, it is difficult to generate the perfect candidate set using such queries, and to find all the relevant publications from these queries. The variant publication curator of UniProtKB/Swiss-Prot explained that they are currently curating variant publications that contain any of the following: 1) missense nucleotide substitutions resulting in amino-acid changes, 2) nonsense nucleotide substitutions producing a stop codon and resulting in protein truncation, 3) small in-frame nucleotide insertions resulting in the insertion of few amino-acids at the protein level, and 4) small in-frame nucleotide deletions resulting in the deletion of few amino-acids at the protein level. If a publication does not contain any of the above, it is marked as "TBD" (to be determined) or "not curatable." We used UniProtKB/Swiss-Prot human data downloaded on September 20, 2017 [37].

The NHGRI-EBI GWAS Catalog

The GWAS Catalog [3] provides manually curated genome-wide association study information from published results. The Catalog contains more than 50,000 unique SNP-trait associations from over 3,000 published research studies. The GWAS Catalog has strict guidelines for study eligibility. For example, the publications must be array-based genome-wide association studies of humans and examine >100,000 SNPs selected to tag variation across the genome. The detailed criteria of eligible publications for the GWAS Catalog are explained on their website [19], and if publications do not meet these criteria, they are not curated. All the publications need to be determined as eligible by at least two curators to be included in the knowledge base. For the triage process, the curators use general query terms to find GWAS-related publications in PubMed and manually filter the papers that do not meet the criteria. We downloaded the October 11th 2017 issue of the GWAS Catalog from its website.

Dataset construction

We collected publications from UniProtKB/Swiss-Prot [1, 2], the GWAS Catalog [3], and mycoSet [20]. Since we did not have curated gold-standard negative data from the knowledge bases, we generated them de novo: we considered an article negative if it was not curated in the knowledge bases, but has one or more variant mentions (found by tmVar). From all the PubMed abstracts and PMC full-text articles with open access, we collected more than 442,000 articles containing variants using tmVar [8, 31].

Using these publications as a gold-standard, we trained classifiers for each knowledge base. We downloaded only PMIDs from each data source. Given a PMID, we used the NCBI Entrez Programming Utilities [38] to collect its title and abstract, journal information, and publication type.

Based on the date (Entrez Date (EDAT)) of each publication, we organized the publications in UniProtKB/Swiss-Prot and the GWAS Catalog into the following three groups: (1) the Before2017 group, which contains articles published before 2017, (2) the JanJul2017 group, which consists of articles published from January 2017 to July 2017, and (3) the AfterAug2017 group, containing articles published after August 2017. We used the papers in the Before2017 group of each knowledge base to design, train, and evaluate the performance of our method in binary classification and used papers in the JanJul2017 group for evaluating the ranking function of the method for curation. We divided the data in this way because we can train only on the past data to predict newly published data in a real-world setting. This data separation setting was used for simulating the real-world triage process. For the Before2017 group of each knowledge base, we randomly collected the same number of negative publications from the negative dataset that are excluded in each positive dataset.

mycoSORT and its dataset—MycoSet

Almeida et al. [20] constructed a dataset and used machine learning methods for a triage task. Their dataset called mycoSet is manually curated and contains publications on enzyme family information. It contains a total of 7,583 PMIDs, of which 9.88% are positive. Their machine learning method used not only the abstracts and titles of the publications, but also used the additional enzyme information (Enzyme Commission numbers and the "RegistryNumber") tagged to the publications. They also performed feature extraction using handcrafted rules, and preprocessed the text using another text mining system called mycoMINE. Naïve Bayes, Logistic model trees, and support vector machine (SVM) were used to classify the publications. In [20], a total of 108 classification results of mycoSORT are listed (three different machine learning classifiers using a total of 36 different ratios of positive and negative publications in

the training set), and the logistic model tree classifier with a negative-positive ratio of 4:1 achieved the best F1 score of 0.575.

Convolutional neural networks (CNN)

We used convolutional neural networks as our machine learning classifiers. CNN is a deep learning method that uses feed-forward multi-layer neural networks such as fully-connected layers, pooling layers and convolutional layers with shared weights for entire inputs [39]. Although CNNs were typically used for image-related works in previous research [40–42], they have recently achieved good results in text related works [22–24, 27, 43]. In addition, it does not require labor intensive feature engineering by domain experts. Hence, our CNN-based approach can be applied to datasets from different databases. We trained three different CNN classifiers on three different sets of PMIDs collected from UniProtKB/Swiss-Prot, the GWAS Catalog and mycoSet, respectively. The details of the datasets are provided in Table 4. The positive to negative ratio of the publications is balanced in all the training, testing and validation sets of UniProtKB/Swiss-Prot and the GWAS Catalog by generating the same amount of negative data.

We built the CNN text classification codes using Keras [44] and TensorFlow [45], based on the implementation [22] available at https://github.com/yoonkim/CNN_sentence. We used the 200-dimensional word2vec vectors of Pyysalo et al. [46–48], which were pre-trained on all the PubMed abstracts and PubMed Central open access full texts, available at http://bio.nlplab.org/. The windows of the filters were set to 3, 4, and 5. A dropout rate of 0.5, a learning-rate of $0.5^{*}10^{-5}$, and a mini-batch size of 50 were used. We followed all the other detailed options as Kim suggested. We used the same parameter settings for all the three classifiers, and we used GeForce GTX 1080 Ti GPU on Linux CentOS 7.4 to train the models. Instead of intensive parameter settings, we used most of the same parameters used in the previous implementation. Our modified source codes and the detailed settings are available at https://github.com/ncbi-nlp/VarTriage.

Assessment methods

We used precision, recall and F1-score to evaluate our classification methods. The scores are calculated as follows:

$$\text{Precision} = (\text{True Positive})/(\text{True Positive} + \text{False Positive})$$

$$\text{Recall} = (\text{True Positive})/(\text{True Positive} + \text{False Negative})$$

$$\text{F1} - \text{score} = (2 \times \text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$$

Table 4. Statistics of the datasets.

	UniProtKB/Swiss-Prot	The NHGRI-EBI GWAS Catalog	mycoSet Positive	mycoSet Negative
Version	Sep. 20, 2017	Oct. 11, 2017 Ver 1.0.1 Studies	-	-
Total # of PMIDs	12,779	3,164	749	6,902
PMIDs with abstracts	11,978	3,143	746	6,575
EDAT before 2017	**11,955**	**2,892**	N/A	N/A
EDAT from Jan to July 2017	**23**	**225**	N/A	N/A

https://doi.org/10.1371/journal.pcbi.1006390.t004

We randomly divided each dataset into 10 folds and used 8 of them for training, 1 for validation and 1 for obtaining the final results which are provided in Table 1. The test sets are used only to obtain the final scores; they are not used for setting parameters or selecting models.

We also used receiver operating characteristic (ROC) curves to evaluate the ranking performance of our methods. ROC curves are plotted using true positive rate (Y axis) and false positive rate (X axis). When the true-positive publications obtain higher scores and true-negative publications obtain lower scores, the area under curve (AUC) becomes larger, which means that the method is ranking the items accurately.

Supporting information

S1 Text. PubMed queries for the query-based methods, and the URLs for the data downloads.
(DOCX)

Acknowledgments

We thank Susan Kim for suggestions and editing of the manuscript. We would like to thank Anthony Rios for his assistance with the deep learning-based text classification.

Author Contributions

Conceptualization: Zhiyong Lu.

Data curation: Maria Livia Famiglietti, Aoife McMahon, Jacqueline Ann Langdon MacArthur, Sylvain Poux, Lionel Breuza, Alan Bridge.

Formal analysis: Kyubum Lee, Chih-Hsuan Wei.

Investigation: Kyubum Lee, Zhiyong Lu.

Methodology: Kyubum Lee, Zhiyong Lu.

Project administration: Zhiyong Lu.

Resources: Maria Livia Famiglietti, Aoife McMahon, Chih-Hsuan Wei, Jacqueline Ann Langdon MacArthur, Sylvain Poux, Lionel Breuza, Alan Bridge, Fiona Cunningham, Ioannis Xenarios.

Software: Kyubum Lee, Chih-Hsuan Wei.

Supervision: Fiona Cunningham, Ioannis Xenarios, Zhiyong Lu.

Validation: Kyubum Lee, Maria Livia Famiglietti, Aoife McMahon, Jacqueline Ann Langdon MacArthur.

Writing – original draft: Kyubum Lee.

Writing – review & editing: Aoife McMahon, Jacqueline Ann Langdon MacArthur, Ioannis Xenarios, Zhiyong Lu.

References

1. The UniProt C. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017; 45(D1):D158–D69. Epub 2016/12/03. https://doi.org/10.1093/nar/gkw1099 PMID: 27899622; PubMed Central PMCID: PMCPMC5210571.
2. Famiglietti ML, Estreicher A, Gos A, Bolleman J, Gehant S, Breuza L, et al. Genetic variations and diseases in UniProtKB/Swiss-Prot: the ins and outs of expert manual curation. Hum Mutat. 2014; 35

(8):927–35. Epub 2014/05/23. https://doi.org/10.1002/humu.22594 PMID: 24848695; PubMed Central PMCID: PMCPMC4107114.

3. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017; 45(D1):D896–D901. Epub 2016/12/03. https://doi.org/10.1093/nar/gkw1133 PMID: 27899670; PubMed Central PMCID: PMCPMC5210590.
4. Keseler IM, Skrzypek M, Weerasinghe D, Chen AY, Fulcher C, Li GW, et al. Curation accuracy of model organism databases. Database (Oxford). 2014;2014. Epub 2014/06/14. https://doi.org/10.1093/database/bau058 PMID: 24923819; PubMed Central PMCID: PMCPMC4207230.
5. Baumgartner WA Jr, Cohen KB, Fox LM, Acquaah-Mensah G, Hunter L. Manual curation is not sufficient for annotation of genomic databases. Bioinformatics. 2007; 23(13):i41–i8. https://doi.org/10.1093/bioinformatics/btm229 PMID: 17646325
6. Bourne PE, Lorsch JR, Green ED. Perspective: Sustaining the big-data ecosystem. Nature. 2015; 527 (7576):S16–7. Epub 2015/11/05. https://doi.org/10.1038/527S16a PMID: 26536219.
7. Van Auken K, Fey P, Berardini TZ, Dodson R, Cooper L, Li D, et al. Text mining in the biocuration workflow: applications for literature curation at WormBase, dictyBase and TAIR. Database (Oxford). 2012;2012:bas040. Epub 2012/11/20. https://doi.org/10.1093/database/bas040 PMID: 23160413; PubMed Central PMCID: PMCPMC3500519.
8. Wei CH, Harris BR, Kao HY, Lu Z. tmVar: a text mining approach for extracting sequence variants in biomedical literature. Bioinformatics. 2013; 29(11):1433–9. Epub 2013/04/09. https://doi.org/10.1093/bioinformatics/btt156 PMID: 23564842; PubMed Central PMCID: PMCPMC3661051.
9. Bandrowski AE, Cachat J, Li Y, Muller HM, Sternberg PW, Ciccarese P, et al. A hybrid human and machine resource curation pipeline for the Neuroscience Information Framework. Database (Oxford). 2012; 2012:bas005. Epub 2012/03/22. https://doi.org/10.1093/database/bas005 PMID: 22434839; PubMed Central PMCID: PMCPMC3308161.
10. Hakenberg J, Voronov D, Nguyen VH, Liang S, Anwar S, Lumpkin B, et al. A SNPshot of PubMed to associate genetic variants with drugs, diseases, and adverse reactions. J Biomed Inform. 2012; 45 (5):842–50. Epub 2012/05/09. https://doi.org/10.1016/j.jbi.2012.04.006 PMID: 22564364.
11. Doughty E, Kertesz-Farkas A, Bodenreider O, Thompson G, Adadey A, Peterson T, et al. Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature. Bioinformatics. 2011; 27(3):408–15. https://doi.org/10.1093/bioinformatics/btq667 PMID: 21138947; PubMed Central PMCID: PMCPMC3031038.
12. Burger JD, Doughty E, Khare R, Wei CH, Mishra R, Aberdeen J, et al. Hybrid curation of gene-mutation relations combining automated extraction and crowdsourcing. Database (Oxford). 2014;2014. Epub 2014/09/24. https://doi.org/10.1093/database/bau094 PMID: 25246425; PubMed Central PMCID: PMCPMC4170591.
13. Yepes AJ, Verspoor K. Mutation extraction tools can be combined for robust recognition of genetic variants in the literature. F1000Research. 2014; 3.
14. Verspoor KM, Heo GE, Kang KY, Song M. Establishing a baseline for literature mining human genetic variants and their relationships to disease cohorts. BMC medical informatics and decision making. 2016; 16(1):68.
15. Poux S, Arighi CN, Magrane M, Bateman A, Wei CH, Lu Z, et al. On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. Bioinformatics. 2017; 33(21):3454–60. Epub 2017/10/17. https://doi.org/10.1093/bioinformatics/btx439 PMID: 29036270.
16. Hirschman L, Burns GA, Krallinger M, Arighi C, Cohen KB, Valencia A, et al. Text mining for the biocuration workflow. Database (Oxford). 2012; 2012:bas020. Epub 2012/04/20. https://doi.org/10.1093/database/bas020 PMID: 22513129; PubMed Central PMCID: PMCPMC3328793.
17. Lu Z, Hirschman L. Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II. Database. 2012;2012.
18. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nature Reviews Cancer. 2004; 4(3):177–83. https://doi.org/10.1038/nrc1299 PMID: 14993899
19. The GWAS Catalog—Methods. Available from: https://www.ebi.ac.uk/gwas/docs/methods.
20. Almeida H, Meurs MJ, Kosseim L, Butler G, Tsang A. Machine learning for biomedical literature triage. PLoS One. 2014; 9(12):e115892. Epub 2015/01/01. https://doi.org/10.1371/journal.pone.0115892 PMID: 25551575; PubMed Central PMCID: PMCPMC4281078.
21. Murphy C, Powlowski J, Wu M, Butler G, Tsang A. Curation of characterized glycoside hydrolases of fungal origin. Database. 2011;2011.
22. Kim Y. Convolutional neural networks for sentence classification. arXiv preprint arXiv:14085882. 2014.

23. Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:151003820. 2015.

24. Zhao Z, Yang Z, Luo L, Lin H, Wang J. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. Bioinformatics. 2016; 32(22):3444–53. Epub 2016/07/29. https://doi.org/10.1093/bioinformatics/btw486 PMID: 27466626; PubMed Central PMCID: PMCPMC5181565.

25. Johnson R, Zhang T, editors. Deep Pyramid Convolutional Neural Networks for Text Categorization. Proceedings of ACL; 2017.

26. Lai S, Xu L, Liu K, Zhao J, editors. Recurrent Convolutional Neural Networks for Text Classification. AAAI; 2015.

27. Santos CNd, Xiang B, Zhou B. Classifying relations by ranking with convolutional neural networks. arXiv preprint arXiv:150406580. 2015.

28. Lee K, Kim B, Choi Y, Kim S, Shin W, Lee S, et al. Deep learning of mutation-gene-drug relations from the literature. BMC Bioinformatics. 2018; 19(1):21. Epub 2018/01/26. https://doi.org/10.1186/s12859-018-2029-1 PMID: 29368597.

29. Amato F, Boselli R, Cesarini M, Mercorio F, Mezzanzanica M, Moscato V, et al., editors. Challenge: Processing web texts for classifying job offers. Semantic Computing (ICSC), 2015 IEEE International Conference on; 2015: IEEE.

30. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of machine learning research. 2011; 12(Oct):2825–30.

31. Wei CH, Phan L, Feltz J, Maiti R, Hefferon T, Lu Z. tmVar 2.0: Integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine. Bioinformatics. 2017. Epub 2017/10/03. https://doi.org/10.1093/bioinformatics/btx541 PMID: 28968638.

32. Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res. 2013; 41(Web Server issue):W518–22. Epub 2013/05/25. https://doi.org/10.1093/nar/gkt441 PMID: 23703206; PubMed Central PMCID: PMCPMC3692066.

33. Bozdag E. Bias in algorithmic filtering and personalization. Ethics and information technology. 2013; 15 (3):209–27.

34. Alain G, Bengio Y. Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:161001644. 2016.

35. Shwartz-Ziv R, Tishby N. Opening the black box of deep neural networks via information. arXiv preprint arXiv:170300810. 2017.

36. Poux S, Magrane M, Arighi CN, Bridge A, O'Donovan C, Laiho K, et al. Expert curation in UniProtKB: a case study on dealing with conflicting and erroneous data. Database (Oxford). 2014; 2014:bau016. Epub 2014/03/14. https://doi.org/10.1093/database/bau016 PMID: 24622611; PubMed Central PMCID: PMCPMC3950660.

37. UniProtKB/Swiss-Prot human data Download. Available from: ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions.

38. Information NCfB. Entrez Programming Utilities 2010. Available from: https://eutils.ncbi.nlm.nih.gov/entrez/eutils/.

39. LeCun Y, Bengio Y. Convolutional networks for images, speech, and time series.

40. Ciregan D, Meier U, Schmidhuber J, editors. Multi-column deep neural networks for image classification. Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on; 2012: IEEE.

41. Lawrence S, Giles CL, Tsoi AC, Back AD. Face recognition: A convolutional neural-network approach. IEEE transactions on neural networks. 1997; 8(1):98–113. https://doi.org/10.1109/72.554195 PMID: 18255614

42. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998; 86(11):2278–324.

43. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. Journal of Machine Learning Research. 2011; 12(Aug):2493–537.

44. Chollet F. Keras 2015. Available from: https://keras.io.

45. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:160304467. 2016.

46. Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou S. Distributional Semantics Resources for Biomedical Text Processing. In: Proceedings of LBM 2013; 2013 p 39–44. 2013.

47. Chiu B, Crichton G, Korhonen A, Pyysalo S, editors. How to train good word embeddings for biomedical NLP. Proceedings of the 15th Workshop on Biomedical Natural Language Processing; 2016.

48. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J, editors. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems; 2013.