Accelerating the Discovery of Semantic Associations from Medical Literature: Mining Relations Between Diseases and Symptoms

Alberto Purpura
IBM Research Europe
alp@ibm.com

Francesca Bonin
IBM Research Europe
fbonin@ie.ibm.com

Joao H. Bettencourt-Silva
IBM Research Europe
jbettencourt@ie.ibm.com

Abstract

Medical literature is a vast and constantly expanding source of information about diseases, their diagnoses and treatments. One of the ways to extract insights from this type of data is through mining association rules between such entities. However, existing solutions do not take into account the semantics of sentences from which entity co-occurrences are extracted. We propose a scalable solution for the automated discovery of semantic associations between different entities such as diseases and their symptoms. Our approach employs the UMLS semantic network and a binary relation classification model trained with distant supervision to validate and help ranking the most likely entity associations pairs extracted with frequency-based association rule mining algorithms. We evaluate the proposed system on the task of extracting disease-symptom associations from a collection of over 14M PubMed abstracts and validate our results against a publicly available known list of disease-symptom pairs.

1 Introduction

Scientific literature is a valuable resource for accelerating scientific discovery in several fields, from computer science to physics, and healthcare (Kumar and Tipney, 2014). However, the overwhelming amount of articles that need to be inspected requires extensive computational approaches as well as modeling knowledge in an appropriate machine readable form. Many efforts have been recently tackling the problem of transforming the unstructured knowledge from scientific papers to knowledge graphs that enable the extraction of actionable insights (Hou et al., 2019; Yadav et al., 2020; Park et al., 2021).

Due to the phenomenal growth of PubMed and MedLine publications (Bretonnel and Lawrence, 2008), the medical domain would particularly benefit from the creation of comprehensive knowledge graphs. Extracting relations between entities is particularly valuable in the medical and biomedical domains where scientists need to extract semantic relations between medical concepts, such as protein and protein, gene and protein, drug and drug, and drug and disease. These relations can be extracted from biomedical literature available from various sources and have already been made accessible in different databases such as BioGRID (Stark et al., 2006) or PDID (Wang et al., 2016). The extraction of these associations from biomedical literature, however, is often time consuming and computationally expensive. Hence, these databases become quickly outdated if they are not updated at the same rate as new scientific discoveries are published. In fact, to the best of our knowledge, the usage of fully automated tools for the extraction of information from these sources is very limited.

In this paper, we propose an efficient end-to-end pipeline for the extraction of semantic relations between medical concepts. We evaluate our approach on disease-symptom associations discovery, but it can also be applied to other relation types and use cases. While disease-symptom information is widely published in medical bibliography, mining such information from literature, electronic health records, or even from user generated content, can accelerate the detection of new symptoms, diseases or variants. Early detection is particularly important in public health surveillance, both in detecting new pandemics (as was the case for COVID-19), identifying new symptoms associated with known diseases (also seen in COVID-19), or even for detecting the resurgence of disease outbreaks in certain countries – as was the case for the Ebola outbreak of 2014 in West Africa. Being able to recognize relations between medical concepts also means that biomedical or clinical texts can be automatically processed at scale, resulting in tools to support decision-making, clinical trial screening, and pharmacovigilance (Yadav et al., 2020).

Proceedings of EMNLP 2022 Industry Track, pages 77–89
December 9–11, 2020.

The proposed pipeline operates similarly to a search engine and can therefore return up-to-date information as its database of documents can be updated in near real time. Our main contributions are:

- a scalable and easily updatable Elasticsearch-based solution to store and query annotated medical literature documents;
- an efficient association rules discovery system based on (i) an association rule mining algorithm, (ii) UMLS semantic network information and (iii) a binary relation classification model trained with distant supervision;
- a solution for human-in-the-loop verification and interpretation of the discovered disease-symptoms associations.

In addition, our solution uses open source libraries and models, and does not require expensive annotation efforts.

An Industry Perspective. With the advance of health informatics, several applications are being developed in the healthcare industry. Automated diagnosis applications as well as symptom checkers are widely used and especially important in low-resource countries to ensure remote medical assistance (Morita et al., 2017). Similarly, health insurance providers and public health organizations are increasingly interested in preventative care to detect early disease onset or complications before patients become expensive and risky to treat.

One of the issues for such application is the collection of information for specific populations. The system proposed in this paper, because of its scalability and ability to interrogate more than 14M abstracts, can represent a useful resource that allows near real time searches to be performed for relation discovering.

2 Related Work

Relation extraction from biomedical text, clinical discharge notes and medical articles have been widely investigated. Among the early systems, Chen et al. (2008) proposed a combination of text mining and association rules to determine the association of drugs and diseases. More recently, increasing numbers of machine learning-based approaches have been developed exploiting supervised learning and feature engineering. Many of them have looked at identifying modifiers related to important clinical entities, such as medication features (Jon D. and Min, 2010; Pathak et al., 2015). An interesting relation extraction task was proposed in the 2010 i2b2/VA challenge (Uzuner et al., 2011). Task organizers released an annotated dataset with medical concepts, assertions and relations, and participants where asked to extract both concepts and assertion as well as specific relations between relevant clinical entities in text. All the top-ranked systems used machine learning-based methods with extensive feature engineering. For example, Grouin et al. (2010) proposed a Support Vector Machine (SVM) based system with additional rules to capture linguistic patterns of relations, while De Bruijn et al. (2011) investigated machine learning using large-dimensional features derived from both textual information and other external sources. Differently from those approaches, we focus on disease-symptom associations, which are useful for a wide variety of healthcare applications and our approach does not require costly manual annotations or extensive feature engineering. More recently, deep learning models have also been applied to solve the same task. Shah et al. (2019) introduce a concept association mining framework based on word embeddings learned through neural networks allowing diseases and related symptoms to be visualized in chronological order. Zhang and Lu (2019) have used a semi-supervised approach based on variational autoencoder for biomedical relation extraction where a multi-layer Convolutional Neural Network (CNN) was used together with bidirectional long short-term memory networks (Bi-LSTMs) to encode drug-drug, protein-protein and chemical-protein interactions. Similar tasks were carried out by Yadav et al. (2020) who introduced a multi task learning framework leveraging a structured self-attentive network together with adversarial learning, and their approach covers also the task of medical concept relation extraction. Our approach explores a different solution, combining pattern mining with NLP algorithms and leveraging more than 14M PubMed abstract, to exploit the vastness of publications to its fullest.

3 Proposed System

The proposed system for the discovery of semantic entity associations is divided into three parts: 1) a sentence annotation system, 2) an Elasticsearch [1] index and 3) a querying and filtering component. The latter can be further divided into three units: (i)

[1] https://www.elastic.co.

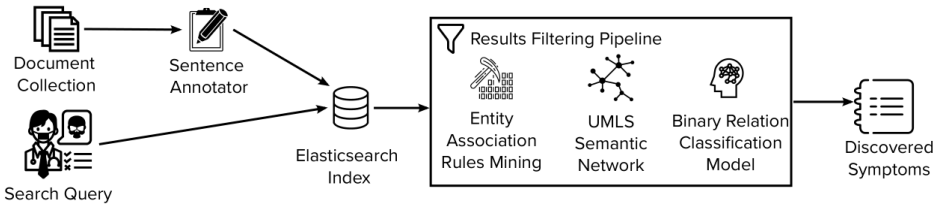

Figure 1: Proposed pipeline architecture.

an Association Rules Mining (ARM) algorithm, (ii) a Unified Medical Language System (UMLS) (Bodenreider, 2004) semantic network-based association rules filter and (iii) a Binary semantic Relation Classification (BRC) model. When deployed, a user can query our system providing a disease name or its UMLS unique identifier and receive a ranked list of symptoms that are likely associated to it, together with a few sentences from the indexed data that motivate the discovered disease-symptom associations. The system facilitates the addition of new documents – such as newly published papers – with the automated annotation of the new textual data and a call to the index API to insert the new records to the index without any updates to the results filtering pipeline. A diagram of the proposed system is depicted in Figure 1.

Documents Annotation and Indexing. When a new document is received, we first split it into sentences and annotate each of them with UMLS entity codes using Scispacy's (Neumann et al., 2019) Natural Language Processing (NLP) pipeline. Scispacy is one of the most efficient and popular NLP annotation libraries for biomedical data and allows for a straightforward integration with the large ecosystem of Python libraries. Its performance is similar to other popular libraries such as MetaMap (Aronson, 2001) but it has a lower inference time (Neumann et al., 2019). For the above reasons, we decide to employ it as the annotation component for our pipeline. Scispacy's entity linker also recognizes mentions of entities from the UMLS database in each of the sentences of the input text. Each of the entities recognized by the annotator has a corresponding UMLS semantic type, such as diseases, symptoms, or anatomy terms, among others. Here, we are going to focus only on the former two entity semantic types – i.e. diseases and symptoms – however, the proposed solution could discover relations between any entity types with minimal adaptations. Once a sentence has been annotated, we store it in an Elasticsearch index. There, we associate different tags to it indicating the UMLS unique identifiers of the entities mentioned in it. We then employ these tags to efficiently retrieve relevant sentences mentioning a certain disease and its symptoms. We use the *keyword* data type from Elasticsearch to store a list of entity identifiers for each sentence record. This strategy allows us to efficiently compute the support (Agrawal and Srikant, 1994) of different diseases and disease-symptom pairs. [2] To summarize, when receiving a new textual document, this component of our pipeline: (1) splits into sentences, (2) adds metadata to each sentence related to the spans of each entity mentioned in it, i.e. a disease, drug or symptom name and their UMLS identifiers, (3) store the sentence and metadata into an Elasticsearch index. Finally, we are aware that the output of this component could contain errors, e.g. mislabeled entities. We protect from them by collecting data from a large number of documents and relying on approaches that are robust to outliers as described in the following sections.

Association Rules Mining (ARM). To discover associations between annotated entity pairs – e.g. a disease and its symptoms – we first query our index to retrieve all sentences mentioning the given disease. For each sentence where the given disease is mentioned, we also obtain a list of other entity identifiers that co-occur with the same disease in that sentence. We then extract all frequent itemsets of size two containing the given disease and a symptom among all possible combinations between any co-occurring pair. For each candidate symptom entity we then compute the support of the itemset of size two containing the given disease by querying the Elasticsearch index. Finally, we compute the confidence of the association rule: $conf(D \Rightarrow C) = supp(D, C)/supp(D)$, where D indicates a disease, C a candidate symptom entity, $supp(\cdot)$ the operation to compute the support of an itemset and $D \Rightarrow C$ an association rule between a disease D and another entity C. We em-

[2] The support of an itemset is defined as the number of times it occurs in a dataset. An itemset could contain a single entity (e.g. a disease) or more than one (e.g. a disease-symptom pair).

ploy this score to rank disease-entity pairs that are most likely to be associated based on the statistical properties of the collected data. This solution for ARM is similar in principle to the popular Pointwise Mutual Information (PMI) (Church and Hanks, 1990) measure of association between words, but allows the discovery of associations between groups of entities larger than two, which can be useful for certain applications.

We also experimented with computing the Lift of each association rule, defined as $Lift(C \Rightarrow D) = conf(D \Rightarrow C)/supp(C)$. However, we found that this metric – which is normalized with respect to the frequency of each symptom – led to lower performances than the former. This is likely due to the fact that it does not take into account the frequency of each candidate symptom alone which is some important information when dealing with noisy annotations coming from text.

Frequency-based measures, however, are not precise enough to recognize all *semantic* disease-symptoms associations in our data. For this reason, we apply two further steps based on the information contained in the UMLS Semantic Network and on our BRC model.

UMLS Semantic Network Based Pruning. From the perspective of our ARM approach, any entity co-occurring with a disease is considered a potential relevant match. To remove association rules involving entities that do not represent a symptom or are not semantically related to a disease, we apply a filtering step based on the information contained in the UMLS semantic network. This is a network of semantic types such as *Disease*, *Symptom* or *Gene*, among others.

Each of these semantic types is associated to a set of UMLS entity identifiers and allows us to identify entities typically recognized as symptoms by the medical community and the relations between them and other entities in the UMLS ontology. In principle, any symptom associated to the *disease* semantic type can be associated to the disease that we are interested in analyzing. The semantic network, however, only yields semantic information between broad semantic concepts that could all be potentially related to each other and does not provide any specific information regarding a particular disease or symptom association. Therefore, this filtering stage allows us to distinguish between a gene or a body part and a symptom that could occur with *any* disease and to remove these *semantically incorrect* associations. For this reason, we employ it in association with the former statistical-based entity strategy and with the BRC model described below to filter out some of the candidate entity pairs.

Binary Relation Classification (BRC). The final step we propose to extract and rank the most likely symptoms for a disease is the BRC model. We randomly sample n sentences (set to 500 in our experiments) where a disease is mentioned together with each of the candidate symptoms and feed them to our BRC model. This model is trained to predict whether two entities in a sentence are semantically related or not. For example, in the sentence: "We conclude that the ability of stress testing to predict <e2>coronary-artery disease</e2> is limited in a heterogeneous population in which the prevalence of disease can be estimated through classification of <e1>chest pain</e1> and the sex of the patient." the two entities *chest pain* (symptom) and *coronary-artery disease* (disease) are semantically related because the sentence is expressing a concept that puts the entities in relation to each other. We are not interested in a specific relation type between the two entities (a symptom-disease relation can be *manifestation-of*, *evaluation-of*, *diagnoses of* according to the UMLS semantic network). Instead, we are interested in recognizing sentences expressing *any* semantic relation between two given entities. The confidence score returned by this component of our pipeline is equal to the ratio of the examined sentences where the same entity pair is classified as semantically related out of the total number of examined sentences where both entities occur (n). For instance, we consider 500 random sentences from our index where a disease like "Asthma" is mentioned together with one of its candidate symptoms such as "Dyspnea". We then classify the relation between these two entities in each of the 500 sentences. We finally compute the BRC association score as the number of sentences where the two entities were classified as related over the total of 500.

The architecture of the proposed model is depicted in Figure 2. We employ a BERT transformer model trained on PubMed abstracts – PubMedBERT (Gu et al., 2021) – available in the Hugging Face library and fine-tune it on the binary classification task on a training dataset we generated automatically with distant supervision. The input to the model is a sentence with entity mentions tags like `<e1>...</e1>` and `<e2>...</e2>` surround-

ing the surface forms of the pair of entities we are interested in. Next, we encode the input sentence with PubMedBERT and then take the representation of the <e1> and <e2> tokens from the last layer of the encoder and concatenate them. This representation of the entities in the input sentence is then fed to a feed-forward neural network with a softmax activation that outputs the probabilities of the two input entities being semantically related to each other given the input sentence. To generate the training data for our model we consider the sentences stored in the aforementioned Elasticsearch index. We select a subset of 300K sentences containing pairs of entities that are related to each other according to the UMLS semantic network – we used 90% of them for training and 10% for validation. These sentences are automatically labeled as containing a relation between a UMLS semantically related entity pair if the same entities are also syntactically related. All other instances from the aforementioned group containing syntactically unrelated entity pairs are considered as negative samples. We say that two entities are *syntactically* related if the root verb of the sentence appears in the shortest path connecting them in the sentence dependency tree. We observed empirically that combining this simple syntactic rule with the information from the UMLS semantic network yields results of a sufficiently high quality to train our BRC model. Other strategies similar to ours are also frequently employed for the creation of relation classification datasets (Smirnova and Cudré-Mauroux, 2018). We fine-tune the transformer model for 10 epochs using Cross-Entropy loss, batch size 64 and learning rate of 2e-5. After training, our model achieved an F1 Score of 0.94 on our randomly sampled validation set.

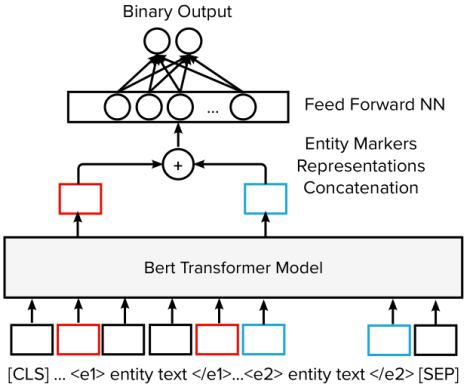

Figure 2: Binary Relation Classification (BRC) model.

Results ranking. The disease-symptoms pairs discovered filtered by the UMLS semantic network are finally ranked in decreasing order of relevance with respect to a combination of the output scores of the ARM and BRC models: Score(d, s) = $\alpha \cdot$ BRC(d, s) + $\beta \cdot$ ARM(d, s) where α and β are parameters optimized using a held out validation set of known entity pairs associations, described in more detail in the next section.

4 Evaluation

We evaluate the proposed system on the task of discovering symptoms associated to different diseases from a collection of over 14M PubMed abstracts published between 2000 and 2022. We assess the performance of the proposed pipeline in terms of Recall, F1 Score, and Precision@k (P@k), with $k \in \{1, 3, 5\}$, defined as the number of relevant items among the top k ranked by the system divided by k. As ground truth data, we consider a subset of the Disease-Symptoms Knowledge Database (Wang et al., 2008) published by Columbia University and available online. [3] This dataset contains a list of disease-symptom pairs extracted from textual discharge summaries of patients at New York Presbyterian Hospital. We only considered 54 out of 134 diseases from the ground truth data since for the remaining 80 no symptoms were ever associated – i.e. mentioned in the same sentence – with the respective disease in our subset of PubMed abstracts. After pruning, our ground truth data contained an average of 1.89 symptoms per disease (min=1, max=5, standard deviation=1.10). For this reason, we limit the number of candidate symptoms returned by our approach to a maximum of 10 per disease after ranking them, as explained before. To determine the parameters α and β used to compute the final ranking score of each disease-symptoms candidate, we perform a 2-fold cross validation optimizing the P@1 performance measure of the entire pipeline.

Disease-Symptoms Associations Discovery. From the evaluation results reported in Table 1, we observe that the proposed pipeline employing both the ARM and BRC models is able to rank among the top 10 symptoms all the ones reported in our ground truth data – i.e. achieves a Recall of 1.00. We also observe that, between the ARM and BRC models, the ARM model is superior in

[3] https://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB.

	P@1	P@3	P@5	Recall	F1 Score
ARM	0.56	**0.33**	**0.25**	0.75	0.29
BRC	0.09	0.10	0.12	0.83	0.21
ARM and BRC	**0.57**	**0.33**	**0.25**	**1.00**	**0.31**

Table 1: Performance evaluation of the proposed pipeline employing only the Association Rules Mining component (ARM), the Binary Relation Classification (BRC) one and both of them (ARM and BRC).

detecting significant associations between diseases and their symptoms in our ground truth. On the other hand, by design, the BRC model prefers symptoms which are mentioned a few times in the corpus but have a very strong semantic association with their respective disease. For instance, the BRC model might more easily detect a rare symptom for a disease which is only mentioned a few times in the scientific literature than the ARM approach. Conversely, the ARM approach distinguishes between frequent and infrequent entities and for this reason is better able to capture associations between common symptoms with their respective diseases. For this reason, the BRC model achieves a lower P@k than ARM but a higher recall since the frequency of co-occurrence of disease-symptoms pairs does not affect the relevance estimation process of this component. Thanks to the different characteristics of the two models, the combined approach employing both elements in the pipeline achieves overall a higher P@k and Recall as the two models are able to complement each other. Overall, the lower P@3 and P@5 values that we observe are motivated by the small number of relevant symptoms for each disease. For example, in most of the cases, a disease has less than 5 recognized symptoms and for this reason its P@5 will be lower than 1.0 even if all the correct symptoms have been retrieved by our model.

Qualitative Evaluation. Quoting Wang et al. (2008) – who created the ground truth of disease-symptom associations that we employ in our evaluation – "One of the limitations in this study is that these associations are based on inpatient reports and therefore may reflect different disease-symptom associations than those that would be acquired using reports from outpatients". For this reason, we include a qualitative assessment of the relevance of the discovered disease-symptoms pairs on our collection of PubMed abstracts. We also use this assessment to show how the support statements provided by our system to each of the disease-symptom association claims could be used in practice to evaluate and interpret the results of our pipeline. As shown in Table 2, for each of the selected disease-symptom pairs, the proposed pipeline also returns at least one sentence that provides some evidence to support its disease-symptom association claims, this allows us to easily verify whether some of disease-symptom pairs are actually not relevant or are just missing from the ground truth. Using this evidence, we manually verified the relevance of each of the top 3 symptoms recognized for each of the diseases by our system. We evaluated the semantic relation of each disease-symptom pair in the top 10 sentences extracted from PubMed abstracts provided as evidence by the model and updated the ground truth if we found any. We share the new disease-symptom associations obtained from this process and the respective PubMed abstracts sentences provided automatically as supplementary material to this paper. As a results of our manual evaluation, we decided to add 70 new disease-symptoms pairs to the ground truth – 1.29 new symptoms for each of the considered diseases. A new disease-symptom association was added to the ground truth if we recognized a statement validating that association among the evidence provided from PubMed by our model. During this process, we observed a few counter-intuitive associations retrieved by proposed pipeline that were semantically correct according to the UMLS semantic network but logically incorrect. For example, we observed 18 associations between different diseases and the UMLS entity "Illness (finding)" which is classified under the semantic type "Sign or Symptom" – e.g. in the sentence "When illness occurs, it is primarily a pneumonic presentation." stating a relation between the disease "Pneumonia" and the "Illness (finding)" entity. We marked these associations as not relevant in our

Disease	Symptom	Support Statement	PMID
Asthma	Dyspnea	Dyspnea is a prominent symptom in asthma.	21635136
Bronchitis	Coughing	For example, midnight worsening of cough is a frequent complaint of patients with laryngitis and bronchitis.	18346860
Dyspnea	Deglutition Disorders	Dysphagia in children most commonly presents as feeding or respiratory difficulty.	14992456

Table 2: Sample of disease-symptom pairs discovered by our approach. We also report one of the statements retrieved in support of each association and indicate the PubMed ID (PMID) of the paper reporting it.

ground truth even if we observed a semantic relation between the entities in the provided sentences. Finally, we repeated the evaluation of the proposed pipeline considering the updated ground truth and observed higher values of P@1, P@3 and P@5 of 0.87, 0.77, 0.47, respectively and a Recall of 0.99.

5 Conclusions, Limitations and Future Work

We describe an end-to-end pipeline for the accelerated discovery of medical entity associations from textual data. We evaluate the proposed approach on the task of extracting disease-symptom pairs from medical literature. The main advantages of the proposed system are (i) its capability to discover new entity associations from collections of millions of scientific abstracts, (ii) the ability to easily include new scientific data in its index and therefore to perform up-to-date predictions, (iii) the independence from human annotations, (iv) the interpretability of its predictions given supporting evidence, (v) its low computational complexity, and (vi) the reliance only on open source libraries and models. We believe the adoption of systems like this in the healthcare industry could help medical professionals and researchers in making better-informed decisions. Such systems could also accelerate scientific discovery by giving researchers the ability to quickly verify potential entity associations claims against scientific literature, or discover new symptoms if used with additional data sources. Our pipeline could also be employed to extend existing resources such as the UMLS ontology with new entity relations.

The proposed approach allows researchers to verify the generated entity associations claims by providing statements from indexed scientific documents that motivate such claims. Despite the encouraging results, the quality of the associations discovered is limited by the accuracy of the entity annotator (incorrect annotations observed in recognizing acronyms may lead to inaccurate entity associations). Similarly, the available UMLS semantic types may not accurately describe the different categories and this can also introduce noise. In addition, the system uses only PubMed abstract, while the entire text could provide more information. As future work, we would like test the pipeline over a larger ground truth, explore relations between other concepts and explore the issue of veracity of the extracted claims in cases where the opinions in scientific literature change over time. We are also planning to evaluate possible improvements to our pipeline, in particular at the stage where the BRC and ARM scores are combined, exploring some machine learning-based approaches for learning to rank.

6 Ethical Considerations

The goal of the proposed system is to support medical professionals and researchers in the accelerated discovery of new entity associations such as new disease-symptom pairs. We show how to do so by relying on the vast collection of medical literature available in PubMed. During the design of the proposed pipeline we paid particular attention to the interpretability and transparency of the results provided by this system. By providing explicit statements in support to each of the discovered associations with references to peer-reviewed scientific publications, we expect users of this system to independently verify the veracity of the provided information and to keep updated the index of publications on which the search system relies. Machine learning models are imperfect and could therefore misinterpret user input or make certain predictions without a having a complete view or understanding of their context. These errors could have serious consequences, especially in the healthcare domain. For this reason, researchers and healthcare professionals employing this system should be aware of possible harms and risks stemming from the use of it, and should implement appropriate safeguards to guarantee the safety of their patients. We also believe that the data we share as supplementary material should not be considered as a verified resource for disease-symptoms associations for any healthcare application as no medical professional reviewed the correctness of our annotations. Finally, since the proposed approach relies on a collection of textual documents for the discovery of new entity associations, it is important that this collection contains an unbiased representation of the entire target population. Otherwise, the system might exhibit a poor performance when interrogated on aspects that might be prevalent in underrepresented populations.

References

Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference of*

Very Large Data Bases, VLDB, volume 1215, pages 487–499. Santiago, Chile.

Alan R. Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Cohen K. Bretonnel and Hunter Lawrence. 2008. Getting started in text mining. *PLoS computational biology*, 4(1):e20.

Elizabeth S. Chen, George Hripcsak, Hua Xu, Marianthi Markatou, and Carol Friedman. 2008. Research paper: Automated acquisition of disease-drug knowledge from biomedical and clinical documents: An initial study. *Journal of American Medical Informatics Assoc.*, 15(1):87–98.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Berry De Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel D. Martin, and Xiao-Dan Zhu. 2011. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association : JAMIA*, 18:557 – 562.

Cyril Grouin, Asma B. Abacha, Delphine Bernhard, Bruno Cartoni, Louise Deléger, Brigitte Grau, Anne-Laure Ligozat, Anne-Lyse Minard, Sophie Rosset, and Pierre Zweigenbaum. 2010. Caramba: Concept, assertion, and relation annotation using machine-learning based approaches.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2019. Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5203–5213, Florence, Italy. Association for Computational Linguistics.

Patrick Jon D. and Li Min. 2010. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of the American Medical Informatics Association : JAMIA*, 17 5:524–7.

Vinod D. Kumar and Hannah J. Tipney. 2014. *Biomedical literature mining*. Springer.

Tomohiro Morita, Abidur Rahman, Takanori Hasegawa, Akihiko Ozaki, and Tetsuya Tanimoto. 2017. The potential possibility of symptom checker. *International Journal of Health Policy Management*, pages 615–616.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327.

Yoonyoung Park, Natasha Mulligan, Martin Gleize, Morten Kristiansen, and Joao H. Bettencourt-Silva. 2021. Discovering associations between social determinants and health outcomes: Merging knowledge graphs from literature and electronic health data. In *AMIA Annual Symposium Proceedings*, volume 2021, page 940. American Medical Informatics Association.

Parth Pathak, Pinal Patel, Vishal Panchal, Sagar Soni, Kinjal Dani, Amrish Patel, and Narayan Choudhary. 2015. ezDI: A supervised NLP system for clinical narrative analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 412–416, Denver, Colorado. Association for Computational Linguistics.

Setu Shah, Xiao Luo, Saravanan Kanakasabai, Ricardo Tuason, and Gregory Klopper. 2019. Neural networks for mining the associations between diseases and symptoms in clinical notes. *Health information science and systems*, 7(1):1–9.

Alisa Smirnova and Philippe Cudré-Mauroux. 2018. Relation extraction using distant supervision: A survey. *ACM Computing Surveys (CSUR)*, 51(5):1–35.

Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. 2006. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl_1):D535–D539.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Chen Wang, Gang Hu, Kui Wang, Michal Brylinski, Lei Xie, and Lukasz Kurgan. 2016. Pdid: database of molecular-level putative protein–drug interactions in the structural human proteome. *Bioinformatics*, 32(4):579–586.

Xiaoyan Wang, Amy Chused, Noémie Elhadad, Carol Friedman, and Marianthi Markatou. 2008. Automated knowledge acquisition from clinical narrative reports. In *AMIA Annual Symposium Proceedings*, volume 2008, page 783. American Medical Informatics Association.

Shweta Yadav, Srivastsa Ramesh, Sriparna Saha, and Asif Ekbal. 2020. Relation extraction from biomedical and clinical text: Unified multitask learning framework. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Yijia Zhang and Zhiyong Lu. 2019. Exploring semi-supervised variational autoencoders for biomedical relation extraction. *Methods*, 166:112–119.

A Supplementary Material

In Table 3, we report all the disease-symptoms pairs that we consider as ground truth for our evaluation in Section 4. We expand the list of associations provided by Wang et al. (2008) by manually evaluating the associations between the top 3 symptoms retrieved for each disease by the proposed pipeline. We considered a disease-symptom association as relevant if we could find at least one sentence – among the top 10 provided by our model from our collection of PubMed abstracts – that confirmed a generalized association between each disease and candidate symptom. For each of the pairs we added, we also report the respective statement that motivated our decision. Disease-symptom pairs for which we do not provide a support statement are part of the associations already provided in (Wang et al., 2008).

Disease	Symptom	Support Sentence (when added to original GT)
Anemia	Fatigue	–
Anemia, Sickle Cell	Chronic pain	Chronic pain affects 50% of adults with sickle cell disease (SCD).
Anemia, Sickle Cell	Syncope	Although benign mechanisms predominate, syncope may be arrhythmic and precede SCD.
Anemia, Sickle Cell	Pain	–
Anxiety state	Fatigue	Fatigue was correlated with depression (r = .40, p < .01), state anxiety (r = .40, p <.01), and trait anxiety (r = .46, p <.01).
Anxiety state	Pain	–
Asthma	Wheezing	This report serves as a reminder to all clinicians that "not all that wheezes is asthma".
Asthma	Dyspnea	Dyspnea is a prominent symptom in asthma.
Asthma	Coughing	–
Bronchitis	Coughing	For example, midnight worsening of cough is a frequent complaint of patients with laryngitis and bronchitis.
Bronchitis	Fever	–
Bronchospasm	Dyspnea	–
Cellulitis	Fever	–
Cellulitis	Pain	–
Cholecystitis	Fever	We report a 55-year-old man who presented with fever and abdominal pain compatible with cholecystitis.
Cholecystitis	Vomiting	We present the case of a 69 year old woman with a history of cholecystitis, who consulted for severe abdominal pain, nausea and vomiting.
Cholecystitis	Abdominal Pain	–
Chronic Obstructive Airway Disease	Signs and Symptoms, Respiratory	COPD is characterized by episodic increases in respiratory symptoms, so-called exacerbations.
Chronic Obstructive Airway Disease	Dyspnea	–
Chronic Obstructive Airway Disease	Coughing	–
Confusion	Seizures	All patients presented with fever and disorientation; 6 of the 9 (66%) presented with seizures.
Confusion	Headache	–
Congestive heart failure	Cheyne-Stokes Respiration	Cheyne-Stokes respiration is frequently observed in congestive heart failure.
Congestive heart failure	Angina Pectoris	He often had episodes of angina at night or during dialysis, and then developed congestive heart failure and was hospitalized.
Congestive heart failure	Dyspnea	–
Deep Vein Thrombosis	Syncope	Symptoms significantly associated with DVT were syncope and chest pain.
Deep Vein Thrombosis	Headache	The clinical picture of deep cerebral vein thromboses (DCVT) usually is acute, combining vigilance disorders, headaches, and focal neurologic deficit.
Deep Vein Thrombosis	Pain	–
Degenerative polyarthritis	Pain	Osteoarthritis is clinically defined mainly by pains upon movement and joint stiffness.
Degenerative polyarthritis	Knee pain	Osteoarthritis (OA) is a major source of knee pain.
Deglutition Disorders	Dyspnea	Dysphagia in children most commonly presents as feeding or respiratory difficulty.
Deglutition Disorders	Hoarseness	–
Dehydration	Diarrhea	–
Dehydration	Vomiting	–

Delirium	Agitation	Agitation can be one of the early signs of delirium or altered mental status (AMS).
Delirium	Malaise	Delirium is highly prevalent in critically ill patients.
Delusions	Psychotic symptom	The psychotic symptoms were variable with delusions and/or hallucinations.
Delusions	Agitation	–
Delusions	Hallucinations, Auditory	–
Diabetic Ketoacidosis	Abdominal Pain	Abdominal pain is a frequent manifestation in patients presenting with Diabetic Ketoacidosis (DKA).
Diabetic Ketoacidosis	Vomiting	–
Diverticulitis	Abdominal Pain	Abdominal pain is the most common complaint in patients with acute diverticulitis.
Diverticulitis	Pain	Acute diverticulitis is a painful disease of the colon characterized by peridiverticular inflammation and/or infection.
Diverticulitis	Fever	–
Epilepsy	Seizures	Epilepsy is the most prevalent neurological disease and is characterized by recurrent seizures.
Epilepsy	Fever	Temporal-parietal-occipital carrefour epilepsy is part of the genetic epilepsy with febrile seizures plus spectrum.
Epilepsy	Headache	Epilepsy bears a bidirectional relationship with headache.
Gastritis	Dyspepsia	Gastritis, GERD, and PUD are the leading causes of dyspepsia.
Gastritis	Diarrhea	Gastritis is an inflammatory disease leading to abdominal pain, nausea, and diarrhea.
Gastritis	Abdominal Pain	–
Gastroenteritis	Diarrhea	Gastroenteritis is a common disease in children, characterized by diarrhea, vomiting, abdominal pain, and fever.
Gastroenteritis	Fever	–
Gastroesophageal reflux disease	Chronic cough	Gastro-esophageal reflux can be the cause of chronic cough.
Gastroesophageal reflux disease	Dyspepsia	Gastritis, GERD, and PUD are the leading causes of dyspepsia.
Gastroesophageal reflux disease	Heartburn	–
Gout	Foot pain	Gout is associated with foot pain, impairment, and disability.
Gout	Fever	In conclusion, gout attacks in elderly patients are associated with fever and higher ESR and CRP levels, often resembling a septic arthritis.
Gout	Pain	–
Heart failure	Dyspnea	–
Hemorrhoids	Pain	–
Hemorrhoids	Diarrhea	–
Hepatitis	Icterus	Onset of hepatitis was defined as jaundice and elevated alanine aminotransaminase (ALT) levels.
Hepatitis	Fever	–
Hypothyroidism	Diarrhea	Diarrhoea and malabsorption are common findings together with hyperthyroidism, whereas constipation is frequently observed in hypothyroidism.
Hypothyroidism	Dry skin	Dry skin may be a manifestation of hypothyroidism.
Hypothyroidism	Fatigue	–
Ileus	Vomiting	Many cases of terminal cancer develop ileus symptoms such as vomiting and abdominal distension.
Ileus	Abdominal Pain	–
Ileus	Constipation	–
Influenza	Headache	–
Mental Depression	Pain	The link between pain and depression lies in the central and peripheral nervous systems.
Mental Depression	Fatigue	Depression was the main factor influencing fatigue among both, MS patients and controls.
Mental Depression	Depressive Symptoms	To study the effect of depression (high levels of depressive symptoms) on social engagement.
Migraine Disorders	Headache	Diet can play an important role in the precipitation of headaches in children and adolescents with migraine.
Migraine Disorders	Pain	Migraineurs have atypical pain processing, increased expectations for pain, and hypervigilance for pain.
Migraine Disorders	Vertigo	Migraine is a common cause of vertigo.
Neuropathy	Neuralgia	Neuropathic pain is the most common type of pain in neuropathy.

Neuropathy	Ataxia	JCV granule cell neuronopathy (JCV-GCN) is caused by infection of cerebellar granule cells, causing ataxia.
Neuropathy	Pain	–
Osteomyelitis	Pain	Osteomyelitis ossis pubis is a painful disorder.
Osteomyelitis	Fever	–
Osteoporosis	Weakness	Osteoporosis is a debilitating disease.
Osteoporosis	Perceived quality of life	These indicate that osteoporosis decreased QOL.
Osteoporosis	Pain	–
Pancreatitis	Pain	Pain is a main complaint of patients with pancreatitis.
Pancreatitis	Icterus	Features of pancreatitis were present in 59, cholangitis in 26 and jaundice in 109 patients.
Pancreatitis	Abdominal Pain	–
Pancytopenia	Hepatosplenomegaly	Examination revealed hepatosplenomegaly associated with pancytopenia.
Paranoia	Sleeplessness	Recent epidemiological studies show a strong association of insomnia and paranoia.
Paranoia	Agitation	–
Parkinson Disease	Motor symptoms	Motor symptoms in Parkinson's disease (PD) patients are usually asymmetric at onset.
Parkinson Disease	Bradykinesia	Motor slowness (bradykinesia) is a core feature of Parkinson's disease (PD).
Parkinson Disease	Tremor	–
Peptic Ulcer	Dyspepsia	Gastritis, GERD, and PUD are the leading causes of dyspepsia.
Peptic Ulcer	Pain	This pain is related to extrahepatic infusion and gastroduodenal ulceration.
Pericardial effusion	Fever	The more common symptoms associated with purulent pericardial effusion are fever, dyspenia, and tachycardia.
Pericardial effusion	Chest Pain	Pericardial effusion was diagnosed because the child suffered chest pain and fatigue.
Pericardial effusion	Dyspnea	–
Pneumonia	Fever	–
Pneumothorax	Respiratory distress	Spontaneous pneumothorax is a recognised cause of respiratory distress in the neonatal period.
Pneumothorax	Hemoptysis	–
Psychotic Disorders	Seizures	Out of these 8 patients, 3 presented with psychosis (12.5%) and 4 (17%) with seizures.
Pulmonary Edema	Respiratory distress	The respiratory distress was initially caused by pulmonary edema and later was caused by severe bronchorrhea.
Pulmonary Edema	Dyspnea	–
Pulmonary Embolism	Syncope	Syncope can be caused by a pulmonary embolism.
Pulmonary Embolism	Dyspnea	–
Pulmonary Embolism	Chest Pain	–
Pyelonephritis	Flank Pain	Symptoms of cystitis are dysuria, frequency, new onset incontinence and malodorous urine while symptoms of pyelonephritis are high grade fever, flank pain and vomiting.
Pyelonephritis	Dysuria	We report the case of a 45-year-old African American man who presented with symptoms of right-sided pyelonephritis, including fever, dysuria, and flank pain.
Pyelonephritis	Fever	–
Respiratory Failure	Dyspnea	–
Thrombocytopenia	Fever	–
Thrombocytopenia	Fatigue	–
Tonic-Clonic Epilepsy	Seizures	Seizures are reported in one quarter, including tonic-clonic, absence, and febrile seizures.
Tonic-Clonic Epilepsy	Fever	Onset in the first year of life by febrile or afebrile clonic and tonic-clonic, generalized, and unilateral seizures, often prolonged, in an apparently normal infant is the first symptom, suggesting the diagnosis.
Tonic-Clonic Epilepsy	Myoclonus	–
Transient Ischemic Attack	Neurologic Symptoms	Transient ischemic attack (TIA) is a cerebrovascular disease with temporary (<24 h) neurological symptoms.
Transient Ischemic Attack	Seizures	Little attention has been paid to the possibility that seizures may be precipitated by TIAs.
Transient Ischemic Attack	Headache	–
Upper Respiratory Infections	Coughing	The commonest form of cough is caused by upper respiratory tract infection and has no benefit to the host.

Upper Respiratory Infections	Fever	–
Urinary tract infection	Fever	Patients' clinical status was dominated by fever due to upper urinary tract infection.
Urinary tract infection	Dysuria	–

Table 3: Extended list of disease-symptoms associations. We provide a supporting statement for each of the associations that we decided to add to the list of disease-symptom associations provided by Wang et al. (2008).