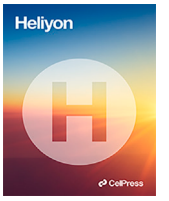
Heliyon 9 (2023) e19265

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Using transfer learning-based causality extraction to mine latent factors for Sjögren's syndrome from biomedical literature

Jack T. VanSchaik [a], Palak Jain [a], Anushri Rajapuri [b], Biju Cheriyan [a], Thankam P. Thyvalikakath [a,b,c], Sunandan Chakraborty [a,*]

[a] Luddy School of Informatics, Computing, and Engineering, Indiana University Indianapolis, Indianapolis, 46202, IN, USA
[b] Indiana University School of Dentistry, Indianapolis, 46202, IN, USA
[c] Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, 46202, IN, USA

ARTICLE INFO

Dataset link: https://github.com/palak-j/Sjogrens_syndrome

Keywords:
Text mining
Causal relationships
Relationship extraction
Sjögren's syndrome

ABSTRACT

Understanding causality is a longstanding goal across many different domains. Different articles, such as those published in medical journals, disseminate newly discovered knowledge that is often causal. In this paper, we use this intuition to build a model that leverages causal relations to unearth factors related to Sjögren's syndrome from biomedical literature. Sjögren's syndrome is an autoimmune disease affecting up to 3.1 million Americans. Due to the uncommon nature of the illness, symptoms across different specialties coupled with common symptoms of other autoimmune conditions such as rheumatoid arthritis, it is difficult for clinicians to diagnose the disease timely. Due to the lack of a dedicated dataset for causal relationships built from biomedical literature, we propose a transfer learning-based approach, where the relationship extraction model is trained on a wide variety of datasets. We conduct an empirical analysis of numerous neural network architectures and data transfer strategies for causal relation extraction. By conducting experiments with various contextual embedding layers and architectural components, we show that an ELECTRA-based sentence-level relation extraction model generalizes better than other architectures across varying web-based sources and annotation strategies. We use this empirical observation to create a pipeline for identifying causal sentences from literature text, extracting the causal relationships from causal sentences, and building a *causal network* consisting of latent factors related to Sjögren's syndrome. We show that our approach can retrieve such factors with high precision and recall values. Comparative experiments show that this approach leads to 25% improvement in retrieval F1-score compared to several state-of-the-art biomedical models, including BioBERT and Gram-CNN. We apply this model to a corpus of research articles related to Sjögren's syndrome collected from PubMed to create a causal network for Sjögren's syndrome. The proposed causal network for Sjögren's syndrome will potentially help clinicians with a holistic knowledge base for faster diagnosis.

1. Introduction

Causal relationships depict important knowledge across many different fields, including medicine and health. Researchers in these fields design and conduct experiments to test causality between two events and publish their findings in research articles. Thus,

* Corresponding author.
E-mail address: sunchak@iu.edu (S. Chakraborty).

https://doi.org/10.1016/j.heliyon.2023.e19265
Received 12 July 2023; Received in revised form 11 August 2023; Accepted 15 August 2023

Available online 22 August 2023

the academic literature records the discovery of new causal relationships or conditions of existing relationships. In this paper, we show how such causal relationships, extracted from the biomedical literature, can help in extracting factors related to a disease. For many diseases, diagnosticians are unaware of *all* factors associated with a disease, which might result in delayed diagnosis. An example of such a disease is Sjögren's syndrome. Sjögren's syndrome is an autoimmune disorder where the immune system destroys glands that produce tears and saliva [1,2] and is also associated with rheumatic disorders [3–5]. The primary symptoms for Sjögren's syndrome are spread across several domain areas, such as dentistry, ophthalmology, and rheumatology. This distribution and the lack of continuity in the communication between dentists and physicians create a critical gap in the proper understanding of the disease's characteristics. Hence, it becomes a challenge for clinicians to timely diagnose Sjögren's syndromein the absence of holistic knowledge. Mining factors from the biomedical text will help create such a holistic knowledge base, allowing clinicians to diagnose such diseases faster. We hypothesize that disease-specific factors can be mined from the biomedical text by extracting causal relationships.

There are numerous way causality can be expressed in natural language text, as a result, extracting causal knowledge from text becomes a challenge. Causality can be stated explicitly (e.g., mosquito bite *causes* malaria) where the relationship is explicitly stated with a clear marker – *causes* [6,7], as well as implicitly (e.g., Last week temperature rose significantly, there were several cases of heat stroke reported), without using causal markers. Due to numerous forms and the presence of implicit causal sentences, extracting causal relationships from text is not trivial. This seriously limits the application of causal relationship extractions from biomedical literature, as given an article, we observe diverse way of expressing causality. This paper proposes a novel method of classifying any given sentence into causal and non-causal sentences. We applied Walsh-Hadamard (WH) Transformation on the input embeddings and added it along the BiLSTM sequence's hidden states in either direction. WH Transform is a non-sinusoidal, orthogonal, and reversible function. It is widely applied in signal and image processing but has not been applied to text data to the best of our knowledge. We compared the performance of the WH-BiLSTM model in classifying causal sentences with a simple BiLSTM baseline model, and our results show that the WH-BiLSTM model's F1 score was 0.91 compared to the baseline model's F1 score of 0.37.

Identifying causal sentences from large documents is the first step toward the holistic extraction of causal knowledge. Past works have used many machine, and deep learning-based approaches [8,9,7,10,30] but they only target explicit causality. Furthermore, they ignore that text presents causality through multi-word expressions or phrases instead of just single words. This paper addresses the root cause of the problem mentioned above; causal relation extraction models are trained on disparate benchmark datasets that vary significantly in lexical composition and annotation style. For example, label sets may vary across data, making certain transferred predictions impossible. In addition, the lack of any causal relationship dataset specifically for biomedical text further limits the application of this technique on such text. To address this, our analysis examines transfer across six unique causal relation datasets that span varying domains, annotation styles, and implicit/explicit causality markups and deploy a novel causal relation extraction model using transfer learning and entity normalization. We fine-tune an ELECTRA-based sentence-level sequence tagging model on causal sentences from several web-based sources. ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) [11] is a BERT variant that uses discriminative pretraining instead of the usual generative pretraining. We train the ELECTRA-based relation extraction model on several datasets and prove its ability to generalize to unseen data, including biomedical text. We apply the fine-tuned model to a set of causal sentences and then use named entity recognition (NER) to identify entities of interest in those phrases. This produces a causal knowledge graph where nodes are entities, and edges are directed causal relationships between those entities.

We conduct an empirical analysis of architectural components, including choice of input embedding, recurrent units, and attention mechanisms. Our analysis showed that the use of attention does not significantly impact model performance in the transfer setting. Thus, to avoid unnecessary parameters and mitigate overfitting, we chose the GRU recurrent unit and forwent the attention layer. Transformer-based embeddings perform well as a contextual word embedding layer in causal relation extraction [12]. We chose ELECTRA due to its superior language understanding capabilities over other transformer models. It also has fewer parameters, reducing the risk of overfitting in the transfer setting. ELECTRA rivals BERT's language understanding abilities with fewer parameters, yet not much work has examined ELECTRA's causal relation abilities.

We create a pipeline that classifies each sentence from a research article into causal and non-causal sentences. We subsequently apply our causal relationship extraction model trained on the six datasets on the causal sentences. We evaluated the performance of our model against several baseline models from previous research and found that our F1 score is consistently better (by ~6%) compared to the other models. We apply this model to a corpus of research article abstracts on Sjögren's syndrome collected from PubMed and manually annotated to identify factors related to Sjögren's syndrome. In this paper, we use Sjögren's syndrome as an example application area. However, our methods are not fine-tuned specifically for Sjögren's syndrome and can potentially be applied to biomedical text on other topics. In this study, we limit our focus only to Sjögren's syndrome and will explore the generalizability of our methods to other topics as part of future work. To evaluate the performance of our model to extract Sjögren's syndrome related factors, we tested the approach on a hand-annotated dataset. The results show that our method has significantly outperformed the baseline models. Finally, we create a causal network using the extracted relationships, and the causal network is shown to reveal new relationships using transitive relationships.

2. Background and related work

Researchers in many fields, design and conduct experiments using methods like, observational studies and randomized control trials to determine whether two events are causally linked, and scholarly articles publish newly discovered causal knowledge emerging from those studies. We see a broad spectrum of work that attempts to retrieve such known causal relationships from a large corpus

of documents and apply them to problems like question answering [13], medical education [14], and financial analytics [15], among others. Expressing causality in a sentence may take several forms. The majority of them are *marked* but maybe *explicit* or *implicit*. Explicit causality has relations that are connected by: (a) causal links (e.g., hence, therefore); (b) causative verbs (e.g., causes, leads to); (c) conditional (e.g., if...then...) [16]. The sentence: "mosquito bites cause malaria," where the word "cause" directly links the cause and effect [6,7] is an example of explicit causality. Implicit causality involves using ambiguous connectives, e.g., *as*, *after* etc., as they are equally likely to be used in causal or non-causal context. For example, "as" is used as a causal marker in the sentence: "There was no debate as the Senate passed the bill on to the House" [6]. Some causal sentences may not have any connectives, for example, the sentence: "Last week temperature rose significantly, there were several cases of heat stroke reported"), where the relationship *rising temperature is the cause of the heatstroke cases* has no causal marker. These are called *unmarked* causal sentences. Causal relationships may span across the sentence. For example, the following two sentences depict a causal relationship [*financial stress → divorce*]: "Being unfaithful can lead to divorce. On the other hand, financial stress is another significant factor." [17].

Past works that addressed this problem can be broadly divided into three groups: rule-based, statistical machine learning (ML)-based, and deep learning-based approaches. Earlier works were primarily rule-based, where linguistic patterns were used to detect explicit causality [18,19]. Girju et al. [20,21] devised a novel approach to a rule-based system, where linguistic patterns were automatically learned instead of manually setting up the rule base. Rule-based methods suffered from a major drawback that it is infeasible to learn all possible rules, and can only extract marked causal sentences, thus leading to poor recall. However, inspired by previous works that used lexico-syntactic patterns to infer causation, a new suite of ML-based methods emerged. The new ML-based methods improved upon the earlier rule-based methods by making the models more generic and not restricted to specific causality patterns. Meuller et al. [22] presented a novel approach and a working prototype that automatically extracts causes and effects, as well as signs, mediators, and conditions, from scientific papers. CausalTriad [23] used a minimally supervised approach, using distributional similarity and discourse connectives. Few other works exploited linguistic structures, such as multi-word expressions [24], N-grams, topics and sentiments [25], lexical patterns [26,20].

With the emergence of deep learning methods, we observe their application in extracting causal relationships from the text. Deep learning methods are capable of learning directly from raw input data without requiring extensive feature engineering, at the same time can handle large-scale datasets effectively. Deep neural networks can efficiently process massive amounts of data. This scalability enables deep learning models to handle complex tasks with enormous amounts of training data, allowing them to generalize well and achieve high performance. As a result, we see more recent works on causality detection from text use deep learning methods [8, 27,28]. Xu et al. [29] used LSTM to learn higher-level semantic and syntactic representations along the shortest dependency path, while Li et al. [30] combined BiLSTM with multi-head self-attention to direct attention to long-range dependencies between words. The latter showed significant improvement when the cause-effect words had a greater separation. Some studies demonstrate that attention, especially of the multi-attention mechanism, shows better performance [30,31]. Zhang et al. [32] combined LSTM with entity position-aware attention to encode both semantic information and global positions of the entities as a result. In recent times we have seen the application of contextual word embeddings and large pre-trained language models in this space. Kyriakakis et al. [10] used BERT [33] and ELMO [34] showed that these models could improve previous state-of-the-art performance with large datasets.

Although RNN-based architectures were producing state-of-the-art performance, some researchers chose to use alternative architectures, such as CNN. An example is by Wang et al. [31], who proposed a multi-level attention-based CNN model to capture entity-specific and relation-specific information and the use of graph-based deep learning models, such as GCN. Zhang et al. [35] proposed a dependency tree-based GCN model to extract relationships that leverages syntactical as well semantical features of the sentences.

SemEval-2010 and ADE datasets are among the most widely used datasets for extracting causality from the text. Many previous works have used the same datasets and developed causal relationship extraction models. These works have used a combination of statistical machine learning and deep learning methods to identify causal relationships from the text. We identified the best-performing models from the literature for each dataset (SemEval and ADE) [17] and compared our performance. Among the best-performing model on SemEval-2010 is a variant of BiLSTM proposed by Li et al. [30]. They combined BiLSTM with multi-head self-attention to direct attention to long-range dependencies between words. Wang et al. [31] also used an attention-based model on CNN instead of BiLSTM. Presently, the best-performing model trained on SemEval-2010 is by Kyriakakis et al. [36]. They used pre-trained language models, such as BERT [33] and ELMO [37] and used Bidirectional GRU with self-ATTention (BIGRUATT) as the base model. Experimental results show that BERT model combined with BIGRUATT performs better on most occasions and scales well with a larger dataset. Among the best-performing models trained on the ADE, the corpus includes the model proposed by Wang and Lu [38], which focuses on jointly modeling entities and relationships. They used a sequence and a table encoder to help each other jointly learn the entities and relations. Zhao et al. [39] used a similar joint modeling technique but proposed Cross-Modal Attention Network (CMAN), has two attention units consisting of BiLSTM-enhanced self-attention (BSA) and BiLSTM-enhanced label-attention (BLA) units.

This study aims to improve upon the limitations and drawbacks of existing methods. One major improvement is that our methods can detect implicit sentences. Our approach do not assume any linguistic structure that expresses causality, thus not dependent on explicit markers. In addition, the proposed WH Transform-supported input embedding helped to identify dependencies that are not detected by many other models. Finally, our model was trained using a diverse dataset that contained many implicit sentences, which improved the visibility of the model, thus making it more generalizable. While our methods supported the detection of implicit sentences, extraction of inter-sentence relationships was not targeted in this study.

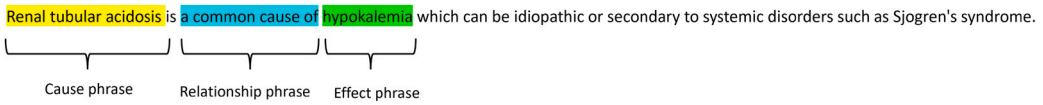


Fig. 1. An example sentence (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3822229/) with a causal relationship that highlights a factor that may lead to Sjögren's syndrome.

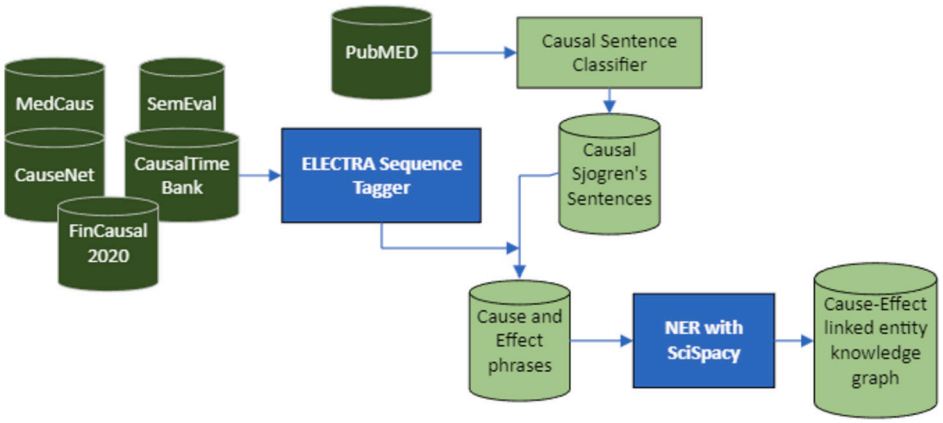


Fig. 2. Overview of the proposed pipeline.

3. Materials and methods

3.1. Problem definition

We define the problem of identifying causal relationships from natural language text as a two-step process - (1) classify any sentence extracted from the research articles as *causal* and *non-causal*, and (2) extract the causal relationships from the *causal* sentences. Fig. 1 shows an example *causal sentence* and the corresponding relationship: ["renal tubular acidosis" *causes* "hypokalemia"]. We define relationship extraction as a sequence tagging problem. In this example, the words 'Renal', 'tubular', and 'acidosis' will have the label *"C"* (cause). The label of 'hypokalemia' will be *"E"* (effect). *"O"* (others) will be assigned to the remaining words, such as 'which', 'can', and 'be'. Fig. 2 presents an overview of the proposed methodology. The conditional probability of our model can be depicted as

$$p(Y|X) = \prod_{i=1}^{n} p(y_i|x_{1\ldots i}, y_{i=1\ldots i-1})$$

3.2. Classification of causal sentences

Recurrent neural architectures, such as LSTMs, BiLSTMs, and GRUs, work well for text classification and entity extraction problems. Even in these models, there is an issue regarding dependencies of words if the distance between them is significant in long sentences. The attention mechanism is typically used for capturing dependencies in long sequences. But the drawback is that it requires a large number of sequential computations. So, we propose to use the Walsh-Hadamard (WH) transformation of input embedding to get the context of the whole sentence. WH transformation is a non-sinusoidal, orthogonal, and reversible function where an input signal is decomposed into a set of basis functions called Walsh functions. It is described by the following binary matrix: $H = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$, a 2×2 WH transform Y of the vector $X \in \mathbb{R}^2$ is $Y = WX = HX$. In general, the WH transform $Y = W_k X$ of a vector $X \in \mathbb{R}^m$ where $m = 2^k$, $k \in \mathbb{N}$ can be expressed via the orthogonal Walsh matrix $W_k \in \mathbb{R}^{m \times m}$ which is generated using the Hadamard matrix and can be recursively constructed [40].

Previously, WH transform has been used in signal processing to reduce the complexity of electronic signal [41], image processing [40] to replace complex layers with a simpler version of WH, and it has been used in Genomics as well [42] in determining different diseases based on DNA and RNA sequencing. These inputs (signals, DNA, RNA) are sequential data similar to text data. So, it is likely that WH transform will have similar advantages for text data. In NLP, for capturing context and dependencies in a long sequence, typically attention mechanism works well but has a complex architecture that ultimately becomes computationally expensive. Hence, we investigate how the context of the whole sentence as input can be provided in a simpler way without using attention mechanism. We used the Walsh-Hadamard transformation to reduce the complexity of this input embedding. To the best of our knowledge, this is the first example of applying WH transform in NLP.

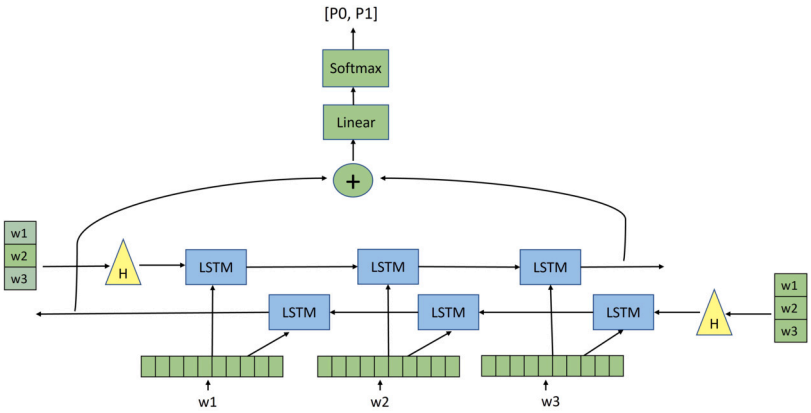


Fig. 3. The overall architecture of the causal sentence classifier. The base is a BiLSTM model with the full sentence matrix going through WH transform and fed into the bidirectional sequence.

The WHT can be used to extract specific features or patterns from the text, allowing for the detection of causal relationships in implicit sentences as well. By analyzing the transformed coefficients, it becomes possible to identify the causal relationship between different parts of the sentences. We combined this method with BiLSTM classification by initializing the hidden state with Walsh Hadamard transformation of whole sentence embedding. Fig. 3 shows the application of WH transform on a base BiLSTM model to build a binary classifier for causal sentences. An alternative method would have been to initialize the hidden state with WH transformation of whole sentence embedding and apply WH transformation to intermediate outputs between all LSTM nodes using the previous node's output. The second approach will require more computational resources, and we will explore that as part of future work.

3.3. Causal relationship extraction

Despite recent advances in deep-learning-based NLP, an ongoing challenge in large-scale causal relation extraction from text is limited size and lack of consistency across training datasets. Data used for benchmarking and training vary significantly in domain and annotation style. Just a fraction of sentences in the widely used SemEval 2010 benchmark dataset [43] are biomedicine related, so using it can not be relied upon to develop a consistent relationship extractor from BioMedical literature. The more recent MedCaus dataset is built from "biomedical" Wikipedia articles [44], but annotations in the dataset span most of the sentences, meaning a model trained on MedCaus alone would fail to isolate biomedical entities of interest. Very little annotated training data is available for causal relation extraction specifically from biomedical literature [45].

To overcome the compounding issues of data availability and annotation discrepancies, we deploy a novel causal relation extraction model using transfer learning and entity normalization. For our approach, we fine-tune an ELECTRA-based sentence-level sequence tagging model on causal sentences from several web-based sources. ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) [11] is a pre-training method for NLP models. It aims to improve the efficiency and effectiveness of language models by focusing on the task of predicting replaced tokens rather than predicting each token in a sequence. The architecture and training method of Electra offers several advantages over traditional pre-training approaches.

In terms of architecture, Electra employs a generator-discriminator framework. The generator is a standard transformer-based language model that predicts each token in a sequence, while the discriminator is another transformer model that tries to distinguish between the original tokens and replaced tokens generated by the generator. This adversarial setup encourages the generator to produce realistic replacements that are challenging for the discriminator to identify. By focusing on the replacement task, Electra enables more efficient training as it does not require the model to predict every token in the input sequence.

The training method of Electra involves two phases: pre-training and fine-tuning. During pre-training, the generator is trained on a large corpus of unlabeled text by applying a masking strategy similar to BERT (Bidirectional Encoder Representations from Transformers). However, instead of predicting the masked tokens, the generator is trained to predict whether the tokens have been replaced or not. In our case, in the fine-tuning phase, the generator is further trained on labeled task-specific data. We added a classification head on top of the Electra model. This head is an additional dense layer followed by a softmax, such as text classification or named entity recognition.

This fine-tuning process allowed the model to adapt to our specific downstream tasks, which identified cause and effect phrases from the sentences. In our case, fine-tuning was done via causal relationship extraction task using our combined labeled dataset (Section 3.4). Causal relationship extraction is classifying each token in the sentence into classes, such as "cause", "effect", or "none" phrases. We anonymized target phrase entities in a sentence using the pre-defined tags such as ##CAUSE## or ##EFFECT##. For example, the sentence "prolonged smoking will lead to COPD", will be represented as "prolonged ##CAUSE## will lead to ##EFFECT##" and the model is trained to predict the entities for each label in the sentence. Through this fine-tuning step our goal is to improve the capability of causal understanding within the underlying model.

Table 1
List of training datasets used. Sentences vary in size, the composition of implicit sentences, and the annotation style.

Dataset	Sentences	Implicit	Domain	Mean tokens per C/E
MedCaus [44]	8682	17%	Medical	8.41 / 7.68
CauseNet-noncause [46]	5000	0%	General	1.61 / 1.5
CauseNet-cause [46]	5000	0%	General	1.53 / 1.46
SemEval 2010	1003 [43]	34%	General	1.06 / 1.02
CausalTimeBank [47]	298	54.7%	News	1 / 0.99
FinCausal2020 [48]	1719	78.7%	Financial	23.72 / 10.26
Total Train	15191			
Total Test	6511			

This training method of Electra has shown improved performance and faster training compared to previous approaches, making it a compelling choice for our application. ELECTRA rivals BERT's language understanding abilities with fewer parameters, yet not much work has examined ELECTRA's causal relation abilities. We train the ELECTRA-based relation extraction model on several source datasets and prove its ability to generalize to unseen data, including biomedical data. We apply the fine-tuned model to a set of causal sentences, extracting cause and effect phrases from those sentences. We then use named entity recognition (NER) to identify entities of interest in those phrases. The above process produces a causal knowledge graph where nodes are entities, and edges are directed causal relationships between those entities.

3.4. Training datasets

We gathered a variety of publicly available, sentence-level, annotated causal relation extraction datasets. These data span several sizes and annotation strategies. To train the model on both explicit and (harder to detect) implicit causality, we included data with various implicit-to-explicit compositions, ranging from entirely explicit (CauseNet) to mostly implicit (FinCausal2020, at 78.7% implicit). Furthermore, tagging schemes tend to be inconsistent across datasets. Some datasets tag single word tokens as cause/effect entities, while others might tag phrases, or even entire sentences, as cause/effect entities. We include datasets with multiple tagging schemes so that the trained model does not solely latch on to longer annotation phrases which tend to be less informative when extracting relationships between entities. The final combined dataset is larger (at 15,191 training sentences) and more diverse than any causal relation extraction dataset that we are aware of. These datasets are briefly described in Table 1.

3.4.1. MedCaus

MedCaus is a dataset consisting of causal sentences mined from "medical articles" in Wikipedia that matched specific seed patterns. While we found that many sentences in this dataset are medical or biological, some general sentences (E.g., "The eastern water is saltier because of its proximity to Mediterranean Water") seem to be captured as well, so we have labeled them as a "General" domain dataset.

3.4.2. CauseNet

CauseNet is a large graph of explicit causal relations from ClueWeb12 and Wikipedia. The CauseNet graph has a precision subset, which we use as a source of explicit causal sentences. For our purposes, we subsampled a collection of 5,000 sentences that contain the explicit markers "cause", "caused", "causing", etc. (CauseNet-cause). The other subsample of CauseNet we used is a collection of 5,000 sentences that do not contain variants of the "cause" marker (CauseNet-noncause), which contains sentences with explicit causal markers like "leads to", "due to", etc. Some preliminary results indicated that causal relation extraction models trained on CauseNet do not improve beyond data sizes of a few thousand, hence the cap of 5,000.

3.4.3. SemEval 2010 Task 8

SemEval 2010 Task 8 [43] is a multi-way classification dataset. It has widely been used as a general domain benchmark for evaluating relation extraction tasks. Causal relation extraction literature has mainly focused on the Cause-Effect relations in this data which represent 12.4% of the entire dataset. We use only the Cause-Effect relations for our analysis.

3.4.4. Adverse Drug Effect

Adverse Drug Effect (ADE) [49,50] contains sentences explaining the adverse effects of drugs using causal sentences. It has been curated from 1,644 PubMed abstracts and contains 6,821 causal sentences. However, this dataset has minimal variation in terms of syntax and vocabulary, and in all sentences, the causality is expressed through the verb *"causes"* and its variation.

3.4.5. Causal-TimeBank

Causal-TimeBank [47] consists of causal annotations of the TempEval-3 corpus [51], which consists of news articles. We only consider sentence-level relations for uniformity across other datasets, although Causal-TimeBank also contains document-level relations.

ELECTRA cause/effect phrase annotations	"…patients with pSS scored high on neuroticism and anxiety and low on sociability." (Cause Phrase / Effect Phrase)
Named entities from SciSpacy:	"…patients with pSS scored high on neuroticism and anxiety and low on sociability."
Phrases are normalized to overlapping entities:	pSS → sociability

Fig. 4. Example phrase normalization. First, cause and effect phrases are identified with the fine-tuned ELECTRA model. Then the SciSpacy NER model is used to identify "disease" entities. Phrases are normalized to any entities that overlap with the cause or effect phrase, producing the final node used in the graph. ELECTRA identified a part of the word "sociability" as the effect phrase due to the Wordpiece embedding used by ELECTRA, which uses subword segmentation. This issue is resolved through the process of phrase normalization.

3.4.6. FinCausal2020

The FinCausal2020 dataset is a benchmark for detecting and extracting causal relations in financial text. For our purposes, FinCausal was limited to relations contained in single sentences.

3.5. Model training

The best performing sequence labeling-based causal relation extraction models use a three-layer approach: (1) a contextual word embeddings layer, (2) a bidirectional recurrent layer, followed by (3) an attention layer [30]. We conducted experiments that found the particular choice of recurrent unit (i.e., LSTM vs. GRU) and the use of attention does not significantly impact model performance in the transfer setting. Thus, to avoid unnecessary parameters and mitigate overfitting, we choose the GRU recurrent unit and forgo the attention layer. Transformer-based embeddings perform well as a contextual word embedding layer in causal relation extraction [12]. We chose ELECTRA due to its superior language understanding capabilities over other transformer models. It also has fewer parameters, reducing the risk of overfitting in the transfer setting.

Sentences from the data sources described in Table 1 were combined, shuffled, and randomly assigned to a 70%-30% train-validation split. We took a sequence labeling approach to relation extraction, as this allows for the most compatibility across the various training datasets. We could simplify all token labels to one of either "O", "C", or "E" (other, cause, and effect). Contiguous output labels were combined to predict a single label. Model architecture consisted of an ELECTRA tokenization and embedding layer, which created contextual embeddings of each input token via a forward pass of the ELECTRA model. This was followed by a BiGRU recurrent layer with hidden and output states of size 256 that were concatenated to size 512. A linear layer was used atop the output embedding layer, with an input size of 512 output size of 3 for each label. We used a softmax loss function as in our case the downstream task is a multi-label classification task.

Hyperparameters used in training the ELECTRA-based sequence tagger are as follows: Minibatch size was 16; Number of output labels was 3; Maximum sequence length (number of tokens) was selected to be 256, which accommodated all the training data and fit in the ELECTRA-Small model. We found 10 training epochs to be sufficient in terms of loss minimization. An ADAM optimizer was used with standard $\beta = (0.9, 0.999), \epsilon = 1e-8$. The learning rate of 5e-5 was determined empirically. A linear layer was used atop the output embedding layer, with an input size equal to the embedding dimension (512 in our case) and an output size equal to the number of labels. A softmax loss function was used.

3.6. Phrase normalization

The datasets used to train the ELECTRA model varied in annotation style. For example, in CauseNet, causes and effects may be labeled as a single word or token, while Medcaus' annotations are typically longer phrases spanning several tokens. The variability in training annotation length meant that the ELECTRA model's predictions also varied in length. However, a helpful knowledge graph should have normalized entities as nodes. To normalize the predicted cause and effect phrases, we used a pretrained scientific Named Entity Recognition (NER) model under the SciSpacy framework. Notably, we used the en_ner_bc5cdr_md model, trained on the BC5CDR corpus, which has disease and chemical entity labels. Within each sentence, we identified named entities that overlapped with the cause-and-effect phrases predicted by our ELECTRA model. This produced normalized cause and effect for each cause and effect phrase. This process is outlined in Fig. 4.

4. Results

We evaluate this work in two phases - (Task 1) evaluate the performance of the causal relationship extraction model, and (Task 2) validate the findings on a large biomedical literature dataset using transfer learning. In this paper, we validate our findings on a set of research articles related to Sjögren's syndrome. We will explore the generalizability of our approach by replicating our methods on other topics as part of future work.

4.1. Performance of the causal relationship extraction model

This task is further sub-divided into two phases - evaluation of (1) the causal sentence classification, and (2) extraction of the causal relationships from the causal sentences.

Table 2
Classification results for the causal sentence classifier with a baseline comparison. Precision-recall is shown separately for each class - "0": non-causal, "1": causal and average F1 score is shown.

Model	Precision	Recall	F1 score (avg)
BiLSTM	0.51(0) 0.65(1)	0.98(0) 0.04(1)	0.37
BiLSTM with WH Transform	0.87(0) 0.95(1)	0.94(0) 0.88(1)	0.91

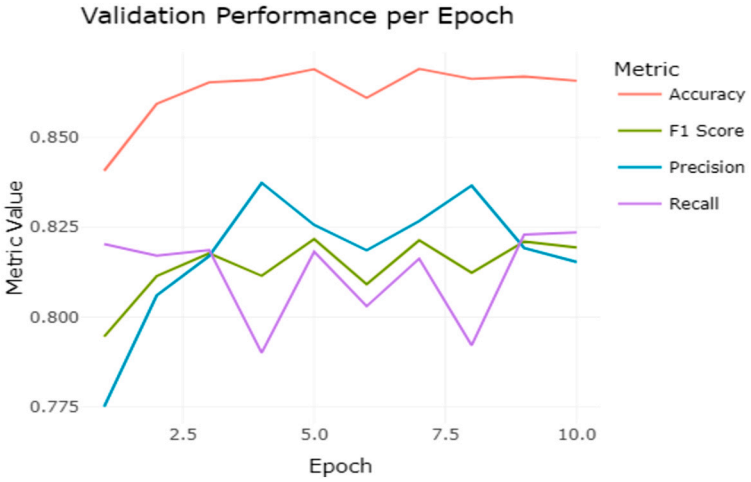


Fig. 5. Performance on all datasets combined over epochs.

4.1.1. Causal sentence classifier

We implemented a BiLSTM-based model with the Walsh-Hadamard (WH) value of the input sentence to identify causal sentences. We randomly sampled 20,000 sentences from the combined dataset (Table 1) dataset for training the model and used another 2,000 for testing. The experimental results are shown in Table 2. The BiLSTM model with the WH values outperformed a baseline model, where the baseline model was the same BiLSTM network without the WH-transformed input of the entire sentence. Given that the only difference in this model is the WH transformed embeddings of the entire sentence, we can infer that by initializing the hidden state with WH value, we are giving the context of the whole sentence together, which can be helpful for the classification of sentences. This eliminates the need for computationally expensive attention weights to *remember* long dependencies. However, to strengthen this claim, we need more experimental results with WH values as inputs in other models to conclude that the WH values have merit in the classification task. As this is not directly related to extracting factors for Sjögren's syndrome, we will explore this as part of future work.

4.1.2. Causal relationship extraction

We applied the final ELECTRA model trained on the combined dataset on each dataset separately to observe the model's transferability. We envision that a model trained on a wide variety of datasets, encompassing different domains, lexical compositions, and annotation styles, will help build a more transferable and generalizable model. By testing, using the test set (not seen during training) part of each dataset, we could measure under what conditions and annotation styles the model performed well. Such a model will also be more suitable for general relationship mining from biomedical literature text. Fig. 5 shows the performance variation over epoch count across all the datasets combined. Although the accuracy value has consistently increased with epoch count, other metrics have shown some variation. For example, we see the precision reaches its peak at epoch 4 and then drop down again to reach the same value (0.837) at epoch 8. Recall shows an opposite trend, where it falls down to 0.790 and 0.792 at epochs 4 and 8, respectively, before reaching the maximum value at epoch 10. Given that the average Performance across all metrics was best at epoch 10, we decided to use this version for all our analyses. The Performance might improve by increasing the number of epochs, but we completed the training at 10 for all the experiments reported in this paper. This choice is mainly due to the time taken and the resources required to continue the training process.

The Performance of this model on each of the datasets (Section 3.4) is shown in Table 3. The precision, recall, and F1 score for extracting causal relationships have been more than or close to 90%. The lowest Performance was on the CausalTimeBank, with an F1 score of 0.84. One plausible explanation is that this dataset had the least number of sentences (298) compared to other datasets with more than 1,000 sentences (ref: Table 1). Lesser number of sentences mean less variation in the data. As a result, the model might underfit with respect to the data and might not generalize well to unseen examples.

Many previous works have used the SemEval-2010 and ADE datasets to develop causal relationship extraction models. We compared our results on these datasets with the best-performing models from the literature [17]. While we have discussed these

Table 3
Performance of the final ELECTRA model on each of the individual training datasets.

Dataset	Precision	Recall	F1
SemEval	0.841	0.943	0.886
ADE	0.883	0.847	0.864
CausalTimeBank	0.807	0.884	0.842
CauseNet_cause	0.929	0.955	0.942
CauseNet_noncause	0.929	0.952	0.941
MedCaus	0.924	0.924	0.924

Table 4
Comparison of performance with selected related works.

Dataset	Model	F1 Score
SemEval-2010	Li et al. [30]	84.6
	Wang et al. [31]	88.0
	Kyriakakis et al. [36]	**90.6**
	Our approach	88.6
ADE	Gurulingappa et al. [49]	70.0
	Wang and Lu [38]	80.1
	Zhao et al. [39]	81.1
	Our approach	**86.4**

works in detail in Section 2, we present our findings from this comparative analysis in Table 4. Our model outperformed other top models trained on ADE. On the other hand, for SemEval-2010, our model was marginally poorer than Kyriakakis et al. [36]. Considering the performance across datasets, our model is likely to perform at par or better than other models. The improved performance of our model can be attributed to the fact that our model has *seen* much more variations in terms of causal sentence type and how the relationship tokens (i.e., causes and effects) are annotated. As a result, our model generalizes better compared to those models. In addition, using ELECTRA as the base model helped in extracting these relationships better, as ELECTRA has demonstrated superior language understanding capabilities [11].

4.2. Latent factor identification from biomedical literature

There are no dedicated datasets for causality extraction based on biomedical literature text. Thus, it is challenging to mine causal relationships without a contextual training dataset. We used the model trained on several datasets across different domains to show how transfer learning can help alleviate this problem. We show that our model can be applied to a new dataset for causal relationship extraction without the need to retrain the model on the new dataset. In this second phase of evaluation, we apply the causal relationship extraction model trained on the six datasets directly on a corpus of biomedical literature (Sjögren's syndrome dataset) (Section 4.2.1) to identify causal sentences and the corresponding cause-and-effect phrases to extract factors related to Sjögren's syndrome.

4.2.1. Dataset

A basic PubMed search was used to produce an initial corpus of text related to Sjogren's syndrome. The search returned 2,350 abstracts comprising 26,000 unique sentences. Some rule-based filtering was applied to these sentences to retain sentences that contained the term Sjögren's syndrome or one of its variants (e.g., "SS", "pSS"). This filtering process ensured that the relationships we extract will provide information about Sjögren's syndrome. After applying the causal sentence classifier (Section 3.2), we identified 5,656 sentences from the abstracts as having at least a 90% probability of containing causal relationships.

4.2.2. Findings

We extract the causal relationships from this text and claim that the opposite label (either *cause* or *effect*) when the term "Sjögren's syndrome" or its variants is detected as *cause* or *effect*, to be the factor related to Sjögren's syndrome. We present a set of selected causal-effect pairs extracted through our model in Table 5. In these examples, we see that Sjögren's syndrome can appear as a cause as well as an effect, which represents the possibility of how factors associated with Sjögren's syndrome are mentioned in the text and the capability of our method to detect them. In these selected examples, we see different factors, such as signs and symptoms (e.g., "loss of secretion", "xerophthalmia") and associated conditions (e.g., "annular erythema", "non-Hodgkin's lymphoma").

We validated our findings with a manually annotated dataset with ground truth labels. We selected a set of 1,058 sentences for annotation, and two annotators with a background in health informatics and experience in Sjögren's syndrome research were asked to annotate the sentences. The annotators labeled relevant factors of Sjögren's syndrome from those sentences. The details of the annotation process is described in Appendix A. Then we used precision, recall, and F1 score to compare the Performance of our approach with several baseline models using the ground truth labels. Table 6 summarizes the findings. The baseline models were Named Entity Recognizers (NER), and some of them, such as BioBERT and Gram-CNN, were pre-trained on biomedical text.

Table 5
Selected examples of extracting factors by mining causal relationships.

	Sentence	Cause	Effect
1	Hypokalemic paralysis is a rare presentation of Fanconi syndrome (FS) caused by Sjogren's Syndrome.	Sjogren's Syndrome	Hypokalemic paralysis
2	Primary Sjogren's syndrome (pSS) is a chronic systemic autoimmune disease that leads to sicca symptoms, mainly xerophthalmia and xerostomia.	Primary Sjogren's syndrome	sicca symptoms, mainly xerophthalmia and xerostomia
3	sjogrens syndrome (SjS) is an autoimmune condition that primarily affects salivary and lacrimal glands, causing loss of secretion.	Sjogren's syndrome	loss of secretion
4	71-year-old woman in whom the diagnosis of possible causes of the development of annular erythema, led the team to identify primary Sjogren's syndrome (SS).	development of annular erythema	primary Sjogren's syndrome
5	Primary Sjogren's syndrome (pSS) is characterized by lymphocytic infiltration of the exocrine glands resulting in decreased saliva and tear production.	Primary Sjogrens Syndrome	decreased saliva and tear production
6	Development of non-Hodgkin's lymphoma (NHL) is the major adverse outcome of Sjogren's syndrome affecting both morbidity and mortality.	Sjogren's syndrome	non-Hodgkin's lymphoma
7	Enthesis zones are important in the formation of pain in the musculoskeletal system in SS patients	Enthesis zones	SS patients
8	Some studies have reported that anti-moesin antibodies have been detected in autoimmune diseases with which SS is closely associated.	anti-moesin antibodies	autoimmune diseases with which SS is closely associated
9	Sjogren's syndrome was suspected based on edentulous state in a middle-aged woman with multisystem involvement	edentulous state	Sjogren's syndrome
10	Autoimmune workup showed antinuclear antibodies with a titer of 1:400 and positive anti SSA (Ro) antibodies that led to the diagnosis of Sjogren's syndrome.	antinuclear antibodies	Sjogren's syndrome

Table 6
Comparative performance.

Model	Precision	Recall	F1-score
Bi LSTM	0.45	0.84	0.59
Glove Embeddings + CNN	0.47	0.72	0.56
Bi LSTM + CRF	0.05	0.4	0.1
BioWordVec + CNN [52,53]	0.48	0.74	0.58
BioBERT [54]	0.39	0.55	0.46
Gram-CNN [55]	0.52	0.74	0.61
Our approach	**0.89**	**0.84**	**0.86**

The results (Table 6) show the central hypothesis of this work that causal relations can be used to extract certain factors associated with Sjögren's syndrome holds. Retrieval performance is better than the baseline methods, but on many occasions, factors are present in a sentence without any causal semantics.

4.3. Causal network from biomedical literature

We created a causal network by combining the individual cause-effect pairs. In this network, each cause-effect pairs were represented as two nodes connected by a directed edge from cause to effect. Then the nodes were merged based on similarity (i.e., same names) to have connected components combining the initially isolated pairs. This network provides additional information through new components, such as a chain of transitive causal relationships, mediators, and confounders of existing relationships through triangular structures. Through this causal network, we have observed that *Tubulointerstitial nephritis* is the most common renal disease caused by Sjögren's syndrome and may lead to *renal tubular acidosis (RTA)*, which in turn may cause *osteomalacia* [56,57]. Even though the entire sequence chain was not directly observed in the data we used, the network could weave the individual relationships and create a more holistic view of the knowledge. Fig. 6 presents a part of the network.

4.3.1. Evaluation of the causal network

The final knowledge graph describes causal relationships between diseases, conditions, and symptoms of Sjogren's Syndrome. However, no comparable knowledge graph exists by which we can evaluate our final product. Thus, we conducted a manual analysis to conduct the evaluated graph. Of the 1,229 edges in the final knowledge graph, 500 were randomly chosen as a more reasonably

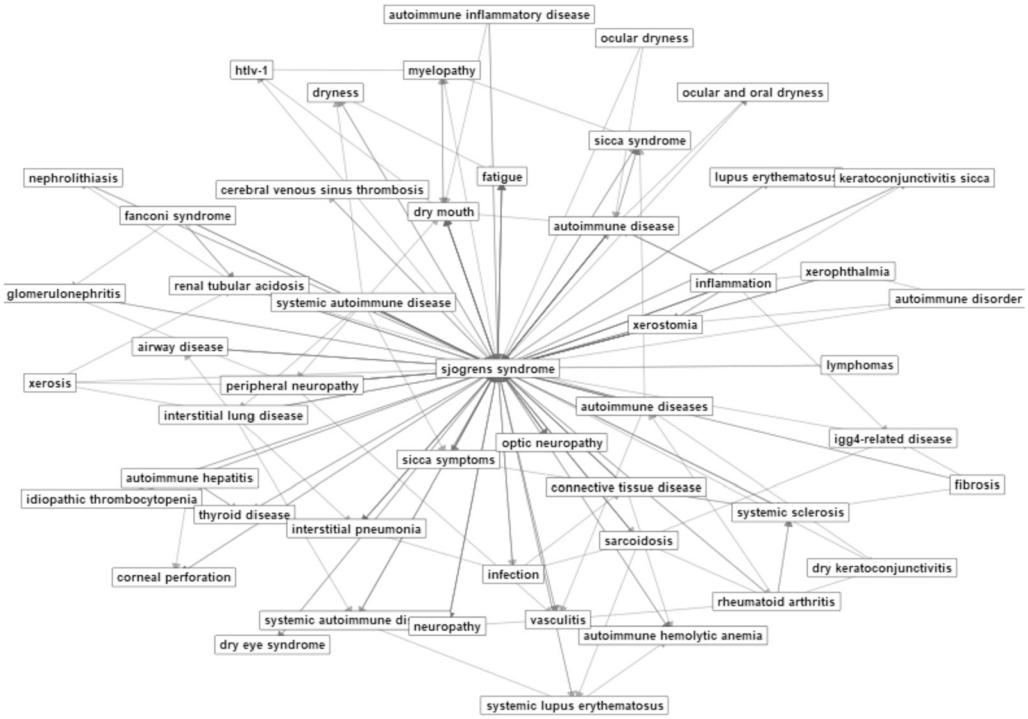


Fig. 6. Nodes in the knowledge graph induced by the above method contain too many nodes for a reasonable static visualization. To view important nodes in the graph, we calculated the eigenvector centrality of each node, then pruned the graph to the nodes with the highest 50 eigenvector centralities. The result is shown above.

sized subsample for manual review. The 500 subsampled edges were again randomly ordered for each of the two distinct reviewers. Reviewers were presented with the cause and effect entities of each edge as well as the source sentence. Reviewers were instructed to label each edge as either having a relationship between entities or not. As measured by the F1 score, the agreement between reviewers was 0.823. Since we have no way of determining false negatives, we must rely on precision as an accuracy metric for our graph. According to reviewer 1, the precision was 75.2%, and for reviewer 2, it was 94.4%.

Of the relationships identified, it was not always clear if the relationship was strictly causal (i.e., could be formulated via contrapositive). This is due to the presence of correlative relationships in the training data, but additionally, sentence-level annotation extraction may lack the context to make such a determination. Thus, we further labeled the true positives in the subsampled edges as either "strictly causal" or "associative". Of the 376 true positives, 169 (44.9%) were able to be identified as strictly causal.

We were interested in the subgraph of annotated relationships that were identified as strictly causal. In the practical setting, such graphs could be helpful to clinicians for diagnosis or prognosis or researchers for literature review and hypothesis development. In the strictly causal Sjogren's Syndrome subgraph, clear clusters formed, as shown in Fig. 7. A cluster of nephrological signs and symptoms is highlighted. This shows a connection between Sjögren's syndrome and nephrology via renal tubular acidosis.

5. Discussion

One of the long-term goals of this work is to create a nearly exhaustive list of factors about a disease by mining information from the biomedical literature. In this paper, we investigate how a causal relationship extraction model can help to work towards that goal. The factors associated with a disease can be categorized into four classes – "signs and symptoms", "risk factors", "associated conditions" and "diagnostic tests". These classifications were provided by clinicians with experience in diagnosing and treating patients with Sjögren's syndrome. Examples of these labels as annotated by the experts are shown in Appendix Table A.7. As we are using causality to mine information from text, we are likely to extract factors that are either "signs and symptoms" or "associated condition", as these two factors are usually causally related to the disease. To illustrate the application of our work, we chose Sjögren's syndrome as an example and validated our results on a corpus of research articles related to Sjögren's syndrome.

Although we assume that causal relationships can be a useful tool to retrieve disease factors, the present version of the causality extraction tool has some limitations. It assumes that there is only one relationship pair in the sentence. In reality, the sentences, particularly in scientific articles, are much more complex, and one single sentence may have multiple relationships in multiple formats – triangular, i.e., two causes leading to one effect or same cause leading to two effects, transitive relations, and presence of conditions that deems the relationship true. For example, the sentence "**sjogrens syndrome (SS)** is a rare condition characterized by **structural damage and secretory dysfunction of the lacrimal and salivary glands** that leads to **dryness, particularly xerophthalmia**

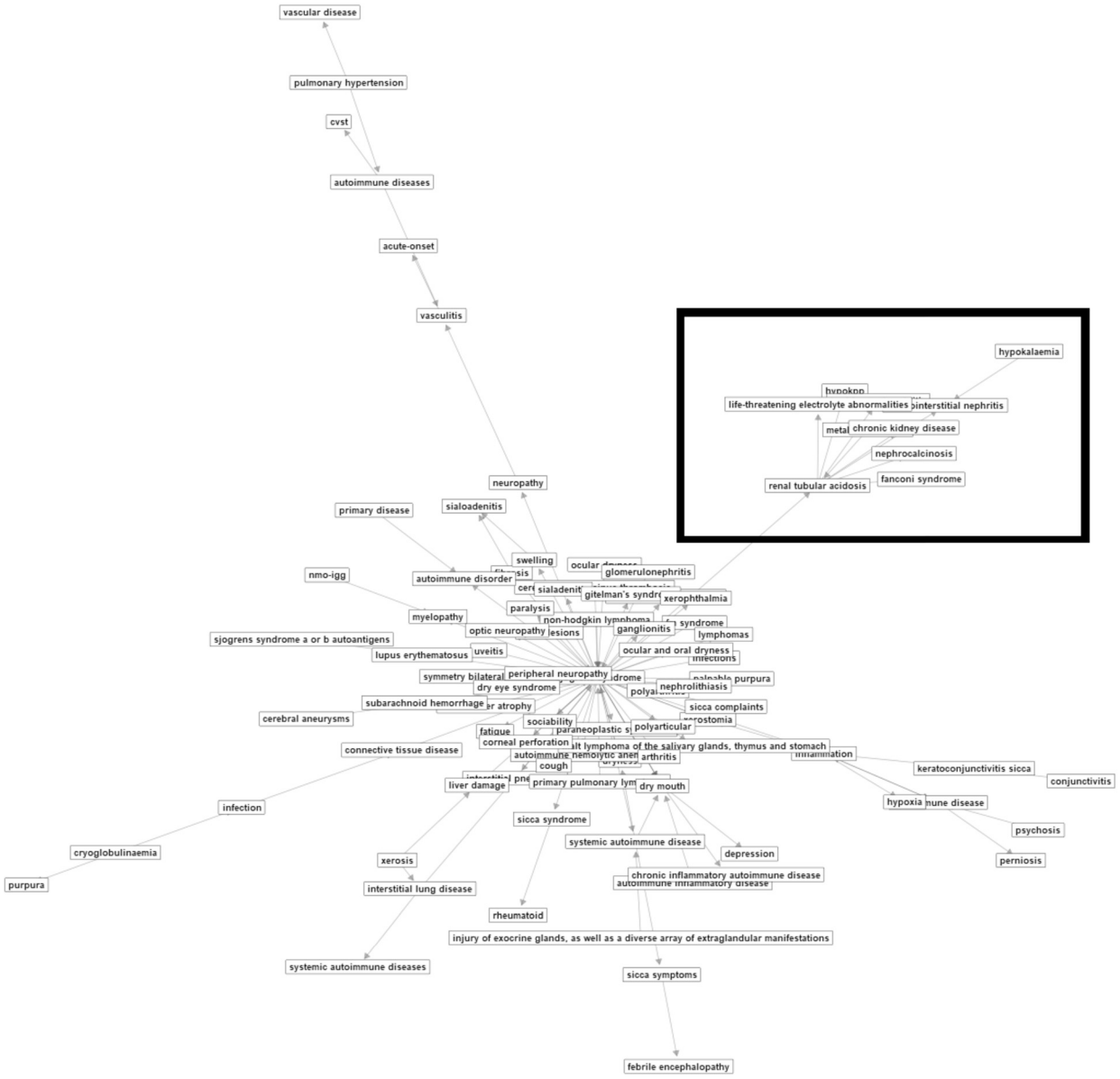


Fig. 7. A subset of the network with edges that were annotated as strictly causal. A cluster of nephrological conditions is highlighted.

(eyes) and xerostomia (mouth)."[1] demonstrates a transitive relation and "**Sjogren's syndrome (SS)** is an autoimmune disease, among the most common ones, that targets mainly the **exocrine glands** as well as **extra-glandular epithelial tissues.**"[2] has a triangular relation, where one event (Sjögren's syndrome) is causing two conditions. This work has not addressed identifying such relations from a single sentence. As part of future work, we will address these limitations and build a more generic causal relationship extraction model that can extract multiple relationships from a single sentence, if present, furthermore, target inter-sentence causal relationships.

The results (Table 6) show the central hypothesis of this work that causal relations can be used to extract certain factors associated with a disease (Sjögren's syndrome in this case) holds. It can retrieve several more factors from the article text compared to other baseline methods. Still, on many occasions, *associated factors* or *signs and symptoms* are present in a sentence without any causal semantics. To achieve the long-term goals and improve the recall of the model, it is essential to identify other relations that bind these factors with the disease. For example, the sentence "Two years after the presentation the patient developed **dyspnea cough and xerostomia**" contains *symptoms*, but due to the absence of a causal semantic, our present model will add this to the list of false

[1] https://pubmed.ncbi.nlm.nih.gov/28862467/.
[2] https://pubmed.ncbi.nlm.nih.gov/29881381/.

negatives. Similarly, our assumption that the disease name (e.g., "Sjögren's syndrome") will be present in the sentence and be part of the cause-effect pair may fail, e.g., the above sentence will not trigger any retrieval by our method. This rationale for using other relations in the future will also help extract other types of factors. As part of future work, we will investigate the relations that will help to discover those factors.

From the clinical perspective, the results emerging from this study will potentially have an important impact. Biomedical literature often contains information about potential factors associated with Sjogren's syndrome. These factors are measurable markers or indicators that can be used to identify and diagnose a particular condition. By analyzing the literature, researchers and clinicians can identify novel factors or gain insights into the significance of known factors for Sjogren's syndrome. These factors can then be used to develop diagnostic tests or improve existing diagnostic methods. However, due to the volume of the literature, manually solving this problem is infeasible. Hence, an automatic method, as presented in this study will help to utilize the vast information available in the literature. This study presents a framework to extract information and its ability to detect factors of Sjögren's syndrome from the literature. While this study shows that the above tasks can be done with reasonable accuracy, a qualitative study evaluating the utility of the extracted factors is beyond the scope of this paper. This qualitative evaluation will require a separate and larger study, involving clinicians who have experience diagnosing and treating Sjögren's syndrome patients. This future study will help assess the quality of our framework's findings and help deploy the methods as a tool in clinical settings and have a real-world impact.

6. Conclusion

This paper presents an innovative approach of using causality to extract factors related to a disease from biomedical literature. We train our model on six different causality datasets to show how transfer learning can help detect causal relationships without any annotated, domain-specific dataset. Using causal relationships, we aimed to extract latent factors about Sjögren's syndrome. Overall, our retrieval method has better precision, recall, and F1 score compared to several supervised baseline models.

Although causal relations could effectively identify many factors, several other types of relations bind the factors with a disease. In the future, to improve retrieval performance, we will investigate other relations and build models that can identify and extract these labels from the text. Furthermore, we will improve our causal relationship extraction model to improve the coverage of relationship extraction and be able to extract multiple causal pairs from a single sentence, as well as discover inter-sentence relations.

Funding

This material is based upon work supported in part by the National Science Foundation under Grant No. 1948322.

Additional information

Supplementary content related to this article has been published online at https://palak-j.github.io/IU/sjogrens_syndrome.html.

CRediT authorship contribution statement

Jack T. VanSchaik: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper. **Palak Jain:** Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data. **Anushri Rajapuri:** Contributed reagents, materials, analysis tools or data. **Biju Cheriyan:** Contributed reagents, materials, analysis tools or data. **Thankam Paul Thyvalikakath:** Conceived and designed the experiments; materials, analysis tools or data; Wrote the paper. **Sunandan Chakraborty:** Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Sunandan Chakraborty reports financial support was provided by National Science Foundation.

Data availability statement

Data associated with this study has been deposited at https://github.com/palak-j/Sjogrens_syndrome.

Acknowledgements

This material is based upon work supported in part by the National Science Foundation under Grant No. 1948322.

Appendix A. Data annotation

A.1. Data extraction and preprocessing

We collected around 2,530 abstracts with 25,525 sentences. These abstracts were extracted from the PubMed database using keywords "Sjogren's Syndrome", "Sjogren" from 2016 to December 2020. Duplicates were removed, and the abstracts were downloaded.

Table A.7
Labels with examples.

Sjogren's Syndrome Concepts	Examples of the literal text match from the sentences.
Signs and Symptoms	"xerostomia", "xerophthalmia", "dry eyes", "dry mouth", "joint pain"
Associated Conditions	"Rheumatoid arthritis", "Systemic lupus Erythematosus", "Squamous cell carcinoma," "Hodgkin's lymphoma"
Diagnostic Tests	"Schirmer Test", "Rose Bengal Test", "Abnormal Flow rate", "Scintigram"
Risk Factors	"Women", "Postmenopausal", "Mean age 40", "Rheumatic Disease"

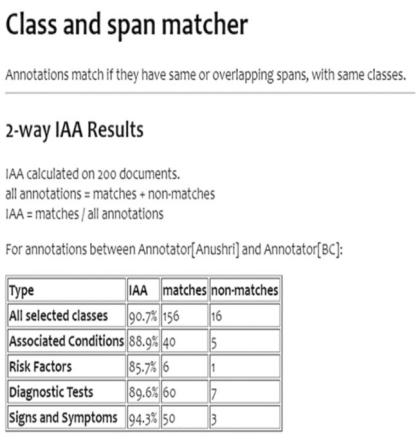
The downloaded data had further additional information such as PMID, Title, Authors, Citation, NIHMS ID, DOI, and abstract text. The abstract text was further cleaned to ASCII text to remove all non-Latin words and letters, and the resulting abstract text was saved to an excel sheet for further usage. Each sentence of the abstract was further broken down and converted into individual text files for annotations. We selected a set of 1,058 sentences for annotation and to be used in all the experiments.

A.2. Annotations guidelines and references standards

We created annotation guidelines for manually annotating Sjogren's Syndrome information that typically dentists seek for their diagnosis of the disease during patient care. We created these guidelines based on the existing literature in dentistry and medicine [3, 4]. Sjogren's related information address by our annotation schema included concepts of Signs and Symptoms, Associated Conditions, Diagnostic Tests, and Risk Factors. Two annotators (A and B) participated in this task and both have advanced knowledge and prior experience with Sjögren's syndrome. We chose the extensible Human Oracle Suite of Tools (eHOST) for this annotation task. Table A.7 summarizes the label and corresponding examples.

A.3. Annotation task

Practice Phase: For this phase, annotators A and B first selected a set of 100 sentences then 501 and lastly 200 from the given dataset and independently annotated them based on the minimal guidelines created. After every set Inter-Annotator Agreements (IAA) were calculated and disagreements between the annotators were resolved through discussion and consensus, and the guidelines were updated subsequently. After this phase concluded, the first author analyzed each annotation set to identify annotation patterns. This cycle continued till a good score of IAA was achieved thus representing an excellent agreement between the two researchers. The analysis results were then discussed among the annotators and served to refine the guidelines.
Adjudication phase: Finally, the final set of annotations were adjudicated and overseen by the annotator C. To create the gold standard to be used on the remaining 2000 annotations. During this phase, annotator C was free and discussed the annotations with the actual annotator to understand his/her reasoning.
Results: After the first set of 100 and 501 sentences, the IAA score was a fair 48.4% and 53.5% with a moderate increase of 5.5%. In discussing the disagreements, the annotators' existing domain knowledge and inference were playing a key role in identifying the concepts. Therefore, for the next set of 200 sentences, a strict ground rule was set, as "The annotations should be text-bound. The annotators domain knowledge and interpretation should play a minimal role in annotation and the annotator should be only concerned with what is explicitly stated in the text. The annotators should also provide basis and justify the annotation and its concept". Following this and the updated guidelines IAA was recorded to be 90.7% (Fig. A.8).


Class and span matcher

Annotations match if they have same or overlapping spans, with same classes.

2-way IAA Results

IAA calculated on 200 documents.
all annotations = matches + non-matches
IAA = matches / all annotations

For annotations between Annotator[Anushri] and Annotator[BC]:

Type	IAA	matches	non-matches
All selected classes	90.7%	156	16
Associated Conditions	88.9%	40	5
Risk Factors	85.7%	6	1
Diagnostic Tests	89.6%	60	7
Signs and Symptoms	94.3%	50	3

Fig. A.8. Screenshot of eHOST tool summarizing the inter-annotator performance and agreement.

References

[1] F.B. Vivino, Sjogren's syndrome: clinical aspects, in: Special Issue: Sjogren's Syndrome, Clin. Immunol. 182 (2017) 48–54.
[2] C.Q. Nguyen, A.B. Peck, Unraveling the pathophysiology of Sjogren syndrome-associated dry eye disease, Ocul. Surf. 7 (1) (2009) 11–27.
[3] C.G. Helmick, D.T. Felson, R.C. Lawrence, S. Gabriel, R. Hirsch, C.K. Kwoh, M.H. Liang, H.M. Kremers, M.D. Mayes, P.A. Merkel, et al., Estimates of the prevalence of arthritis and other rheumatic conditions in the United States: Part I, Arthritis Rheum. 58 (1) (2008) 15–25.
[4] L.E. Brown, M.L. Frits, C.K. Iannaccone, M.E. Weinblatt, N.A. Shadick, K.P. Liao, Clinical characteristics of ra patients with secondary ss and association with joint damage, Rheumatology 54 (5) (2015) 816–820.
[5] L.E. Brown, M.L. Frits, C.K. Iannaccone, M.E. Weinblatt, N.A. Shadick, K.P. Liao, Clinical characteristics of RA patients with secondary SS and association with joint damage, Rheumatology 54 (5) (10 2014) 816–820 [Online]. Available: https://doi.org/10.1093/rheumatology/keu400.
[6] E. Blanco, N. Castell, D.I. Moldovan, Causal relation extraction, in: Lrec, 2008.
[7] A. Ittoo, G. Bouma, Extracting explicit and implicit causal relations from sparse, domain-specific texts, in: International Conference on Application of Natural Language to Information Systems, Springer, 2011, pp. 52–63.
[8] T. Dasgupta, R. Saha, L. Dey, A. Naskar, Automatic extraction of causal relations from text using linguistically informed deep neural networks, in: Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, 2018, pp. 306–316.
[9] Q.-C. Bui, B.Ó. Nualláin, C.A. Boucher, P.M. Sloot, Extracting causal relations on hiv drug resistance from literature, BMC Bioinform. 11 (1) (2010) 101.
[10] M. Kyriakakis, I. Androutsopoulos, A. Saudabayev, et al., Transfer Learning for Causal Sentence Detection, 2019.
[11] K. Clark, M.-T. Luong, Q.V. Le, C.D. Manning, Electra: pre-training text encoders as discriminators rather than generators, arXiv preprint, arXiv:2003.10555, 2020.
[12] V. Khetan, R. Ramnani, M. Anand, S. Sengupta, A.E. Fano, Causal bert: language models for causality detection between events expressed in text, in: Intelligent Computing, Springer, 2022, pp. 965–980.
[13] H.Q. Yu, Dynamic causality knowledge graph generation for supporting the chatbot healthcare system, in: Proceedings of the Future Technologies Conference, Springer, 2020, pp. 30–45.
[14] M.S. Yin, M. Pomarlan, P. Haddawy, M.R. Tabassam, C. Chaimanakarn, N. Srimaneekarn, S.-U. Hassan, Automated extraction of causal relations from text for teaching surgical concepts, in: 2020 IEEE International Conference on Healthcare Informatics (ICHI), IEEE, 2020, pp. 1–3.
[15] D. Chen, Y. Cao, P. Luo, Pairwise causality structure: towards nested causality mining on financial statements, in: CCF International Conference on Natural Language Processing and Chinese Computing, Springer, 2020, pp. 725–737.
[16] C. Khoo, S. Chan, Y. Niu, The many facets of the cause-effect relation, in: The Semantics of Relationships, Springer, 2002, pp. 51–70.
[17] J. Yang, S.C. Han, J. Poon, A survey on extraction of causal relations from natural language text, Knowl. Inf. Syst. (2022) 1–26.
[18] C.S.G. Khoo, S. Chan, Y. Niu, Extracting causal knowledge from a medical database using graphical patterns, in: ACL '00, 2000, pp. 336–343.
[19] D. Garcia, et al., Coatis, an nlp system to locate expressions of actions connected by causality links, in: International Conference on Knowledge Engineering and Knowledge Management, Springer, 1997, pp. 347–352.
[20] R. Girju, D.I. Moldovan, Text mining for causal relations, in: Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference, AAAI Press, 2002, pp. 360–364 [Online]. Available: http://dl.acm.org/citation.cfm?id=646815.708596.
[21] R. Girju, Automatic detection of causal relations for question answering, in: Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering-Volume 12, Association for Computational Linguistics, 2003, pp. 76–83.
[22] R. Mueller, S. Hüttemann, Extracting causal claims from information systems papers with natural language processing for theory ontology learning, in: Hawaii International Conference on System Sciences, 2018.
[23] S. Zhao, M. Jiang, M. Liu, B. Qin, T. Liu, Causaltriad: Toward Pseudo Causal Relation Discovery and Hypotheses Generation from Medical Text Data, 2018.
[24] S. Sasaki, S. Takase, N. Inoue, N. Okazaki, K. Inui, Handling multiword expressions in causality estimation, in: IWCS 2017—12th International Conference on Computational Semantics—Short Papers, 2017.
[25] D. Kang, V. Gangal, A. Lu, Z. Chen, E. Hovy, Detecting and explaining causes from text for a time series event, arXiv preprint, arXiv:1707.08852, 2017.
[26] D. Bollegala, S. Maskell, R. Sloane, J. Hajne, M. Pirmohamed, Causality patterns for detecting adverse drug reactions from social media: text mining approach, JMIR Public Health Surveill. 4 (2) (2018).
[27] J. Chen, Q. Zhang, P. Liu, X. Qiu, X. Huang, Implicit discourse relation detection via a deep architecture with gated relevance network, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, Aug. 2016, pp. 1726–1735 [Online]. Available: https://aclanthology.org/P16-1163.
[28] E.M. Ponti, A. Korhonen, Event-related features in feedforward neural networks contribute to identifying causal relations in discourse, in: Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-Level Semantics, Association for Computational Linguistics, Valencia, Spain, Apr. 2017, pp. 25–30 [Online]. Available: https://aclanthology.org/W17-0903.
[29] Y. Xu, L. Mou, G. Li, Y. Chen, H. Peng, Z. Jin, Classifying relations via long short term memory networks along shortest dependency paths, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1785–1794.
[30] Z. Li, Q. Li, X. Zou, J. Ren, Causality extraction based on self-attentive bilstm-crf with transferred embeddings, Neurocomputing 423 (2021) 207–219.
[31] L. Wang, Z. Cao, G. De Melo, Z. Liu, Relation classification via multi-level attention cnns, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1298–1307.
[32] Y. Zhang, V. Zhong, D. Chen, G. Angeli, C.D. Manning, Position-aware attention and supervised data improve slot filling, in: Conference on Empirical Methods in Natural Language Processing, 2017.
[33] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding, 2018.
[34] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, Jun. 2018, pp. 2227–2237 [Online]. Available: https://aclanthology.org/N18-1202.
[35] Y. Zhang, P. Qi, C.D. Manning, Graph convolution over pruned dependency trees improves relation extraction, arXiv preprint, arXiv:1809.10185, 2018.
[36] M. Kyriakakis, I. Androutsopoulos, A. Saudabayev, J. Ginés i Ametllé, Transfer learning for causal sentence detection, in: Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, Florence, Italy, Aug. 2019, pp. 292–297 [Online]. Available: https://aclanthology.org/W19-5031.
[37] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2018), 2018.
[38] J. Wang, W. Lu, Two are better than one: joint entity and relation extraction with table-sequence encoders, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Nov. 2020, pp. 1706–1721, Online: Association for Computational Linguistics.
[39] S. Zhao, M. Hu, Z. Cai, F. Liu, Modeling dense cross-modal interactions for joint entity-relation extraction, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, International Joint Conferences on Artificial Intelligence Organization, 7 2020, pp. 4032–4038, main track.
[40] H. Pan, D. Badawi, A.E. Cetin, Fast Walsh-Hadamard transform and smooth-thresholding based binary layers in deep neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4650–4659.

[41] A. Jayathilake, A. Perera, M. Chamikara, Discrete Walsh-Hadamard transform in signal processing, IJRIT Int. J. Res. Inf. Technol. 1 (2013) 80–89.
[42] X. Zhao, D. Pompili, Walsh-Hadamard transform of dna methylation profile for the classification of human cancer cells, in: Proceedings of the 5th International Conference on Bioinformatics and Computational Biology, 2017, pp. 26–29.
[43] I. Hendrickx, S.N. Kim, Z. Kozareva, P. Nakov, D.O. Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, S. Szpakowicz, Semeval-2010 task 8: multi-way classification of semantic relations between pairs of nominals, arXiv preprint, arXiv:1911.10422, 2019.
[44] F. Moghimifar, G. Haffari, M. Baktashmotlagh, Domain adaptative causality encoder, arXiv preprint, arXiv:2011.13549, 2020.
[45] A. Akkasi, M.-F. Moens, Causal relationship extraction from biomedical text using deep neural models: a comprehensive survey, J. Biomed. Inform. 119 (2021) 103820.
[46] S. Heindorf, Y. Scholten, H. Wachsmuth, A.-C.N. Ngomo, M. Potthast, Causenet: towards a causality graph extracted from the web, in: CIKM, ACM, 2020.
[47] P. Mirza, R. Sprugnoli, S. Tonelli, M. Speranza, Annotating causality in the tempeval-3 corpus, in: Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL), 2014, pp. 10–19.
[48] D. Mariko, H. Abi-Akl, E. Labidurie, S. Durfort, H. De Mazancourt, M. El-Haj, The financial document causality detection shared task (fincausal 2021), in: Proceedings of the 3rd Financial Narrative Processing Workshop, 2021, pp. 58–60.
[49] H. Gurulingappa, A.M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, L. Toldo, Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports, J. Biomed. Inform. 45 (5) (2012) 885–892.
[50] H. Gurulingappa, A. Mateen-Rajpu, L. Toldo, Extraction of potential adverse drug events from medical case reports, J. Biomed. Semant. 3 (1) (2012) 15.
[51] N. UzZaman, H. Llorens, L. Derczynski, J. Allen, M. Verhagen, J. Pustejovsky, Semeval-2013 task 1: tempeval-3: evaluating time expressions, events, and temporal relations, in: Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), 2013, pp. 1–9.
[52] Y. Zhang, Q. Chen, Z. Yang, H. Lin, Z. Lu, Biowordvec, improving biomedical word embeddings with subword information and mesh, Sci. Data 6 (1) (2019) 1–9.
[53] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, T. Mikolov, Fasttext. zip: compressing text classification models, arXiv preprint, arXiv:1612.03651, 2016.
[54] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.
[55] Q. Zhu, X. Li, A. Conesa, C. Pereira, Gram-cnn: a deep learning approach with local context for named entity recognition in biomedical text, Bioinformatics 34 (9) (2018) 1547–1554.
[56] T. Du, X. Liu, W. Ye, W. Ye, C. Li, Primary Sjögren syndrome-associated acute interstitial nephritis and type 3 renal tubular acidosis in a patient with thin basement membrane nephropathy: a case report, Medicine 99 (32) (2020).
[57] O. Aiyegbusi, L. McGregor, L. McGeoch, D. Kipgen, C.C. Geddes, K.I. Stevens, Renal disease in primary Sjögren's syndrome, Rheumatol. Ther. 8 (1) (2021) 63–80.