Journal of Biomedical Informatics 123 (2021) 103932

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

A practical approach towards causality mining in clinical text using active transfer learning

Musarrat Hussain [a], Fahad Ahmed Satti [a], Jamil Hussain [b], Taqdir Ali [a], Syed Imran Ali [a], Hafiz Syed Muhammad Bilal [a], Gwang Hoon Park [a], Sungyoung Lee [a,*], TaeChoong Chung [a,*]

[a] Department of Computer Science and Engineering, Kyung Hee University Seocheon-dong, Giheung-gu, Yongin-si, Gyeonggi-do 446-701, Republic of Korea
[b] Department of Data Science, Sejong University, Republic of Korea

ARTICLE INFO

Keywords:
Causality mining
Active transfer learning
Clinical text mining
Machine learning

ABSTRACT

Objective: Causality mining is an active research area, which requires the application of state-of-the-art natural language processing techniques. In the healthcare domain, medical experts create clinical text to overcome the limitation of well-defined and schema driven information systems. The objective of this research work is to create a framework, which can convert clinical text into causal knowledge.

Methods: A practical approach based on term expansion, phrase generation, BERT based phrase embedding and semantic matching, semantic enrichment, expert verification, and model evolution has been used to construct a comprehensive causality mining framework. This active transfer learning based framework along with its supplementary services, is able to extract and enrich, causal relationships and their corresponding entities from clinical text.

Results: The multi-model transfer learning technique when applied over multiple iterations, gains substantial performance improvements. We also present a comparative analysis of the presented techniques with their common alternatives, which demonstrate the correctness of our approach and its ability to capture most causal relationships.

Conclusion: The presented framework has provided cutting-edge results in the healthcare domain. However, the framework can be tweaked to provide causality detection in other domains, as well.

Significance: The presented framework is generic enough to be utilized in any domain, healthcare services can gain massive benefits due to the voluminous and various nature of its data. This causal knowledge extraction framework can be used to summarize clinical text, create personas, discover medical knowledge, and provide evidence to clinical decision making.

1. Introduction

Natural language has provided a key cohesive ingredient for pushing the boundaries of technological advances beyond individuals to the 4th industrial revolution. In textual form, it provides a long term, stable, knowledge base, which can be used to preserve knowledge across generations. Digital evolution in the last century has greatly accelerated this preservation process and provided a means to extract meaningful information from texts, which is tedious task for the human beings to process the plethora of textual documents. Natural Language Processing (NLP) is a divergent field, with state-of-the-art research initiatives looking towards resolving the various challenges of automatic information extraction. Foremost, amongst these challenges is the ability to identify the various concepts and their relationship, which form the epitome of the target corpus [1]. For humans and machines, cause-effect represents an essential relation, which provides ample support for the reasoning and decision making process [2]. Automatic causality detection has benefited greatly from numerous dedicated research efforts [3]. As a result, applications such as information retrieval [4], question answering [5], and event reasoning and predictions [6] have gained valuable improvements through the identification of cause-effect relationships. However, challenges such as the dynamicity of syntax and

* Corresponding authors.
E-mail addresses: musarrat.hussain@oslab.khu.ac.kr (M. Hussain), fahad.satti@oslab.khu.ac.kr (F.A. Satti), jamil@sejong.ac.kr (J. Hussain), taqdir.ali@oslab.khu.ac.kr (T. Ali), s.emran.a@oslab.khu.ac.kr (S.I. Ali), bilalrizvi@oslab.khu.ac.kr (H.S.M. Bilal), ghpark@khu.ac.kr (G.H. Park), sylee@oslab.khu.ac.kr (S. Lee), tcchung@khu.ac.kr (T. Chung).

https://doi.org/10.1016/j.jbi.2021.103932
Received 13 July 2020; Received in revised form 29 September 2021; Accepted 4 October 2021
Available online 8 October 2021

semantics, and the evolution of vocabulary have hindered the development and usage of any generic and cross-domain solution [7].

The commonly used approaches for causality detection, fall into two categories: pattern-based traditional rule bases, and machine learning based automatic classification and entity extraction [2,8,9]. Pattern based approaches are based on partial or complete expert intervention for crafting and verifying the conditions based on the syntactic and semantic analysis of the corpus. This approach, requires intensive human effort and lacks cross-domain generalization. Even after utilizing a substantial amount of human time, the extracted rules cannot cover all possible linguistic patterns and are usually not usable beyond the original domain/corpus. Such an approach also suffers from the diversity in linguistic typology, leading to rules formed for a language based on the Subject-Verb-Object (SVO) sentence structure (Such as English, Chinese, French, and others) not being compatible with those based on other structures such as Subject-Object-Verb (SOV) and others [10].

Automatic machine learning based approaches utilize labeled datasets for extracting causality relationships from unseen data and thereby requires less expert intervention, relatively. With this approach, most human time is spent on labeling the data and verifying the results, while providing a reusable model for cross-domain applications. However, any evolution of labels and change in text can render the model unusable. Additionally, machine learning models, are typically independent of the linguistic topology features and can be customized to work on any sentence structure albeit with some effort towards creating and optimizing language vectors, and incorporating natural heuristics derived from syntactically labelled (supervised learning) or a well distributed large corpus (un/semi-supervised learning) [11].

A solution to managing change in the machine learning models and reducing the expert's intervention is available as Transfer Learning, where the machine can learn new tasks by reusing a foundational model, originally employed for a different but related task in another domain [12–14]. Such a cross-domain application may not replicate the original performance benchmarks, and requires some model tuning and tweaking before becoming useful. Model tuning is achieved with the help of a human expert who provides feedback to the machine learning model for improving its learning tasks, a technique more commonly known as active learning [15,16]. To gain benefits of these two approaches active transfer learning is applied to various tasks in diverse domains [17,18], transferring similar models and improving its performance in a single workflow. This performance is mainly improved by enhancing the pre-trained model with few annotated dataset and expert involvement from the new domain.

Causality mining as an application of causality detection is typically based on two tasks, which includes identification of causal triggers, and causal pairs participating in each relationship [8]. Also known as causal connectives; causal triggers are transitive verbs which form a bridge between causality concepts and identify the cause and its effect. Leveraging the sentence structuring in English language[10], typical causality relation identification methodologies, found in research literature, follow the Noun Phrase (NP) - Verb (V) - NP pattern which corresponds to either Cause - Trigger - Effect or Effect - Trigger - Cause forms (<S→NP-Cause, V→ Verb-Trigger, O → NP- Effect>) [19]. Based on this heuristic, Kaplan and Berry-Bogge [20] provided an early model for creating and using handcrafted linguistic template for causality detection. Kalpana Raja et al. [21], built upon the same idea in addition to identifying and organizing a dictionary based on causal trigger keywords, which was then used to define patterns for causality detection. R. Girju et al. [5] refined the process of identifying the causal verbs by utilizing the WordNet dictionary [22]. Bui et al. [23] applied rule based approach for causal relation extraction on HIV drug resistance. Cole et al. [24] utilized a syntactic parser to convert the SVO structures into SVO triples, which were then passed through various rule based filters for causality detection. S. Zhao et al. [8], pointed towards the existence of diversity in the manner each causal trigger expresses causality. However, the syntactic structure of causal sentences and the way the trigger invokes the causality, can provide satisfactory categorization of the causal triggers, enabling smart application of the causality identification filters. Son Doan et al. [9] presented an application of causal mining by marking several verbs and nouns as causal triggers for extracting causal relations from twitter messages. Saud Alashri et al. [25] proposed a snowball strategy, where the authors defined few causal verbs as "seeds" and enlarged the seed list from climate new text by generalizing the seed verbs. Girju and Moldovan [19] proposed a semi-supervised approach towards causality relation identification by using the underlying linguistic patterns of the corpus.

Many other automatic causal pattern identification methodologies have relied on the evolution of machine learning models. In particular, [26] has presented a causal relation extraction model using unsupervised learning to detect the noun phrases corresponding to the subject and object of the sentence. By analysing an unannotated raw corpus and using Expected Maximization(EM) along with a Naive Bayes classifier, the authors were able to precisely identify 81.29% of causal relations.

On the other hand, E. Blanco et al. [27] utilized a supervised learning approach by first annotating ternary instances as being a causal relation or not, and then applied Bagging with C4.5 decision trees to achieve a precision of over 95% in causal relations and above 86% in non causal ones. These and many other machine learning approaches have been comprehensively classified by [3], which indicates a general trend towards the utilizing of the same, as the models become more mature and stable. Of particular interest are the word embedding methods, which due to their requirement of unsupervised data, scalability, and accuracy have piqued the interest of the NLP research community.

Several initiatives have already led to the state-of-the-art results in completing NLP tasks such as sentiment analysis, text classification, topic modeling, and relation extraction [7]. Zeng et al. [28] classified relations in the SemEval Task 8 dataset using deep convolution neural networks (CNNs). Nguyen et al. [29] introduced positional embedding to the input sentence vector in CNNs for improved relation extraction. Silva et al. [7] proposed a deep learning (CNN) based causality extraction methodology that can detect causality along with its direction. The author addressed the causality detection problem as a three class classification problem, where class 1 indicates the annotated pairs has causal relation with direction entity1 → entity2, class 2 implies the causal relation has the direction entity2 → entity1, and class 3 entities are non-causal.

Ning An et al. [30] has utilized a word embedding with cosine similarity based approach, which uses an initial causal seed list to identify the causal relationships as a multi-class (four-class) classification problem. With one-hot encoding the authors, convert the causal verbs in the seed list and the verbs identified in Noun Phrase(NP)-Verb Phrase(VP)-Noun Phrase(NP) ternary(triples) into encoding vectors. These vectors are then converted into Embedding vectors using Continuous Skip-Gram based on a Wikipedia dataset of 3.7 million articles. Finally the encoded vectors are then compared using cosine similarity and the pair with maximum similarity above a pre-defined threshold value of 0.5 are used to classify the causal relationship and evolve the seed list. This method achieved an average F-score of 78.67%. While this methodology presents a significant improvement on previous research initiatives towards causal relationship identification, it suffers from low accuracy, due to its focus on causal verb identification based on a small initial seed list and its limited extension, and classification based, solely on these verbs meanwhile losing context of the causal phrase.

In this paper we present a novel causal relationship identification framework, which outperforms, the existing solutions in the domain of causality mining in clinical text. This framework uses a multi-dimensional approach, which resolves syntactic and semantic matching problems in clinical textual data, providing causal knowledge which is useful to summarize clinical text for quick review, create patient personas for reapplication of medical procedures and predictive analysis, discovering medical knowledge from volumnous data sources, and provide evidence supporting clinical decision making.

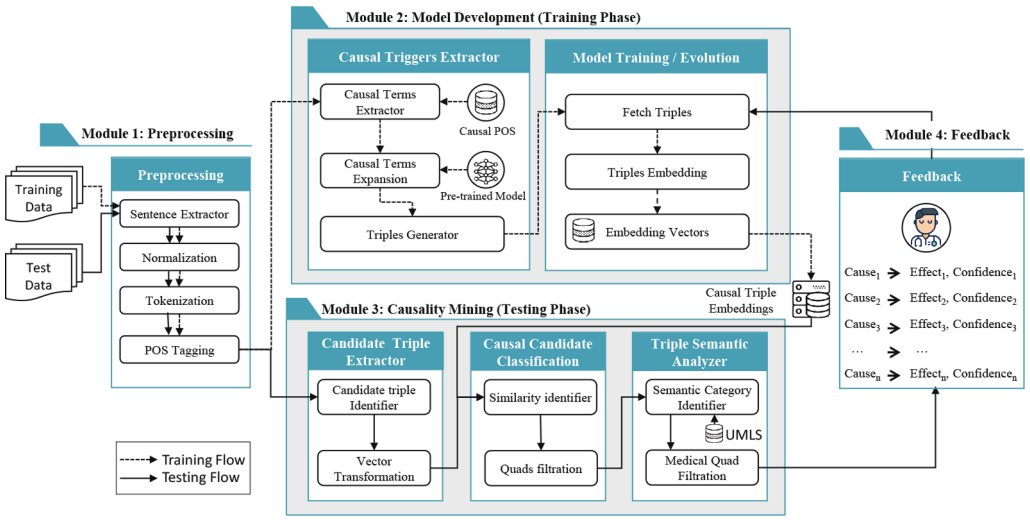

Fig. 1. Proposed methodology workflow.

This novel framework first identifies causal phrases in the form of causal triples (subject, causal verb, and object) using dependency based linguistic patterns[31]. These patterns are extracted from part of speech (POS) tagged sentences in SemEval Task 8 *training* dataset[32]. By removing linguistic elements, which are not used for causality mining, the resultant set of causal triples when compared with the set of unique sentences in the text, is larger in terms of total instances (One sentence can have many causal phrases, the separation of which increases the breadth of our search space) and smaller in terms of words within each instance. The causal triple set maintains the model performance in terms of causality mining, while reducing compute times at subsequent points. Each component of the causal triple is then expanded via transfer learning using pre-trained Google News model [33]. The expanded causal triple in "*NP V NP*" (SVO) form is then converted into embedded vector using Bidirectional Encoder Represenations from Transformers (BERT) [34]. These embedded vectors are then used to calculate a similarity matching score, against embedded causal triples from SemEval Task 8 *test* dataset [32]. The matching scores, and evaluation of the precision-recall curve then provides the matching threshold, over which a triple (and its corresponding phrase) can be classified as causal and under which as non-causal. The embedded vectors from the training dataset and the threshold calculated thus far, are then applied on two test datasets, to classify each test triple as causal or non-casual. The matching confidence score is then used to extend each causal triple, forming the causal quad (subject, causal verb, object, similarity confidence). The noun phrases within these causal quads are then semantically enriched using Unified Medical Language System (UMLS) [35,36], to extend any and all, healthcare terms within these, with their semantic and uniquely identifiable corresponding codes. The extended causal quads are then validated by the expert, producing a list of incorrectly identified phrases. This list is then semantically matched against the trained embedded vectors, and all matches above the threshold are removed. The reduced set of embedded vectors, thus produced, are then re-used for causality detection in the test datasets, thereby completing an active learning loop.

This methodology is further elaborated in Section 2, with experimental setup following in Section 3, the results in Section 4, and the discussion in Section 5. Finally, Section 6 will conclude the paper.

2. Method

Modern medicine and healthcare services have greatly improved the daily human life and yet they are beleaguered by constant evolution of diseases, newfound scientific discoveries, and state-of-the-art engineering inventions. This evolution necessitates the use of information technology in general and natural language processing in particular to mine the plethora of healthcare data, information, and knowledge sources to form computable resources. As a part of this endeavor, we present a framework and its novel application for automatically detecting and classifying causal relationships in healthcare textual data. The framework processes clinical text such as clinical notes and clinical practice guidelines, to extract causal knowledge for enabling the medical experts to perform effective diagnosis, treatment, and follow up.

The framework provide four main service categories/modules; Preprocessing, Model Development (MD), Causality Mining (CM), and Feedback Loop as depicted in Fig. 1. The preprocessing module transforms the input textual corpora into syntactic enriched sentences which are used by both MD and CM modules for training and applying casual relationship identification model, respectively. The MD module extracts causal triples from the annotated dataset and uses various pre-trained models to self-expand and then generate embedding vectors forming the Causal Trigger Trained Model (CTTM). This model is then used to mine candidate causal relations from unseen clinical text by the CM module, subsequently preparing the causal relationships for verification by an expert. A feedback loop based on the experts' assessment towards the correctness of each relationship, is passed to MD for actively improving the CTTM for future applications. Each of these modules is further discussed in the following subsections.

2.1. Preprocessing module

Real world textual data is considered dirty since it contains many defacto linguistic elements which may be a part of daily conversations and routine usage between humans but are not understandable by a computing device. The primary aim of preprocessing is to prepare clinical text for causal phrase extraction which are then used by the MD module to expand the list of causal triggers and by the CM module for semantic comparisons.

The first step of this process is to extract individual sentences from the input corpora using the Natural Language Toolkit (NLTK) [37] sentence tokenizer. Syntactic problems such as redundant text, unrelated information (Explanations, such as this one, in parenthesis which are useful for readers but not required for establishing context), and special characters (−, +, _, etc.) are removed in the normalization step using regular expression. Each processed sentence is then tokenized into words using NLTK word tokenizer. Finally, Part Of Speech (POS) tagging is applied on each word using Standford CoreNLP Parser (version 3.9.2) [38], thereby completing the preprocessing stage. The syntactically

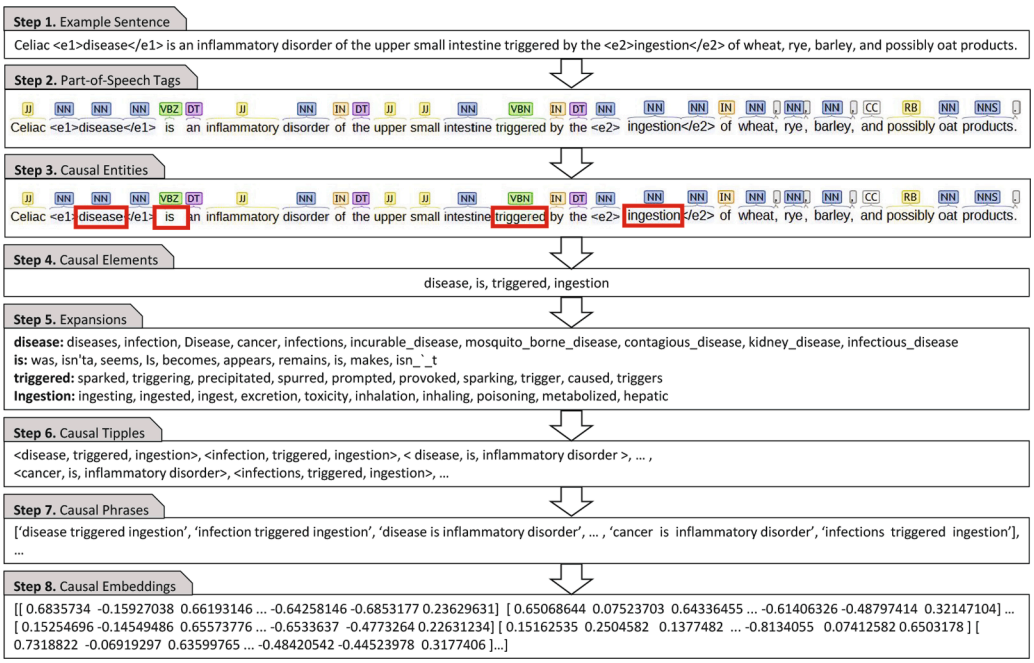

Fig. 2. Training causal trigger extraction example.

enriched sentences are now ready for causal phrase extraction by the MD module and semantic comparisons by the CM modules.

2.2. Model Development (MD) module

The MD module extracts an initial casual trigger list from the syntactically annotated data produced via preprocessing of the training dataset. This list is then expanded using pre-trained models, before being converted into embedded vectors and becoming a part of the CTTM. This process completes in two steps, Causality Trigger Extraction and Model Training/Evolution, which are discussed in the following sub-sections.

2.2.1. Causality trigger extractor

In stage one causal trigger extraction is used to generate a causal triple of the form <NP, VP, NP> which can corresponds to either <Cause, Causal Trigger, Effect> or <Effect, Causal Trigger, Cause>. This process starts by extracting causal triggers which appear as a combination of these noun phrases and verbs from syntactically enriched sentences (while there may be other sentence structures corresponding to causal relationships, in this research we are only focused on processing the aforementioned structures). Since there could be many verbs within each noun, and there can be multiple phrases within each sentence that qualify as a causal triple, we collect the set of all verbs within well-defined noun phrases. We then expand the elements (*NP* and *VP*) of the causal triple using transfer learning technique on a pre-trained model. In the presented approach, we have applied transfer learning using the pre-trained Google News model, which can be replaced with by utilizing other expansion techniques such as, synonym search from WordNet dictionary [22], ConceptNet Numberbatch Model [39], and/or Facebook Fasttext Model [40].

The expansion of each term is restricted to top ten similar words. This choice of selecting only the top ten similar words is driven by the impact of this selection on quantity of operations required for embedding vector generation and their subsequent comparisons.

Once the triples have been expanded, we then apply Cartesian product between the two expanded noun phrases (Expansion set of the 1st and 3rd element of the causal triple) and one of the verb expansion from the causal triple. This increases the number of causal triples, which in turn increases the scope of causal sentences that can be correctly classified in the testing phase.

2.2.2. Model training/evolution

In stage two, the set of causal triples are converted into embedding vectors using pre-trained BERT language models. In order to generate the embedding vectors, the three elements of the causal triple are concatenated by spaces, producing a phrase of the form "*NP V NP*". The collection of these embedded vectors, forms the Causal Triple Trained Model (CTTM). In our experiments, which will be discussed in later sections, we compared 6 BERT Natural Language Inference(NLI) models with mean, max, and cls tokens [41], in terms of their ability to correctly classify causal sentences, from unseen test dataset. Based on the coverage of causal terms by these models, a multi-model approach is well suited for the causality mining task. As a result, each causal phrase is converted into 6 embedding vectors generated via the 6 BERT NLI models. While the space of the CTTM is increased 6-fold, due to this enhancement, it also provides better semantic matching performance, which will be discussed in the results section.

2.2.3. Example of model development (training phase)

An example of this process is shown in Fig. 2. Starting with a sample sentence from our training dataset, which contains the annotated cause and effect entities enclosed within e1 and e2 tags in step 1, we applied preprocessing on it. This produced a POS annotated sentence in step 2, which is used to identify the tagged nouns and verb terms between them in step 3. Each noun term is further expanded to include the preceding adjectives, if any. Any verb terms outside the tagged nouns are ignored. As shown in the step 4 "is" and "triggered" are two of the candidate verbs identified in this process, while "disease" and "ingestion" are their encapsulating noun phrases. In step 5, each of the participating noun and verb phrase is expanded by identifying their closely related alternatives. In step 6, we applied Cartesian product on the sets of two nouns and each verb phrase, producing the set of expanded causal triples. In step 7, the causal triples are converted into causal phrases, which are then converted into embedding vectors as shown in step 8.

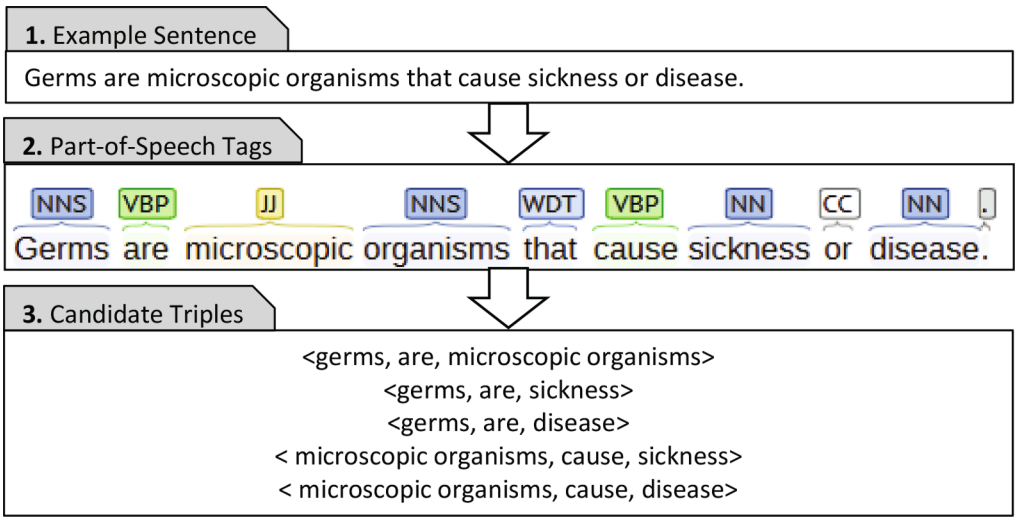

Fig. 3. Test candidate causal triple extraction example.

2.3. Causality Mining (CM) module

The CM module is used for application of the CTTM on unseen, preprocessed test data, for classifying candidate phrases as causal or non-causal. This module utilizes three steps Candidate Triple Extraction, Causal Candidate Classification, and Triple Semantic Analysis, which are described in following sub-sections.

2.3.1. Candidate triple extractor

In the first step, starting with preprocessed sentences from unseen text, the Candidate Triple Extractor, identifies the candidate triples. These candidate triples are obtained by collecting all possible phrases of the form <NP, VP, NP> within each preprocessed sentence. This operation is performed in linear order to collect various candidate causal phrases within each sentence, thus increasing the total number of candidates but greatly reducing the size of individual phrases. For sentences with more than one verb in a sentence, the noun phrases with longer dependencies are discarded. This is to maintain context of the nouns with their nearest verb phrase for matching with our causality identification patterns of SVO. An example of this process is shown in Fig. 3, where the sentence from step 1, is pre-processed in step 2, before candidate triples for the same are generated in step 3. The candidate triples are then converted into candidate phrases ("*NP V NP*"), before the 6 BERT pre-trained models convert each of these into 6 embedded vectors.

2.3.2. Causal candidate classification

Next, we apply the Causal Trigger Trained Model(CTTM) to classify the candidate embedded vectors generated in the previous step, as being causal or non-causal. The CTTM contains embedding vectors for 6 BERT models, which all participate in the causality classification operation, using cosine distance measure to solve this 2-class problem. Each of the BERT model, classifies a candidate triple as causal if the max similarity score is above α_i (where i is the index, corresponding to one of the six models, and α_i is computed using the threshold selection methodology presented in Section 4.1). The causal triple thus identified is expanded by including the similarity score, as a fourth member, thus transforming the triple into a quad of form $< NP, VP, NP, [score_i] >$. Where $score_i$, represents the similarity measure of a participating BERT model. These quads are then filtered using minimum similarity threshold. For 6 BERT models, presented in the Section 4, a candidate triple is thus classified as causal if at least one model classifies it as causal. Additionally, the minimum value of $score_i$, greater than or equal to α_i is retained as the similarity score of the candidate triple. In this way, we can determine the minimum similarity of a candidate triple with most participating models. The final set of quad thus produced, pertains to causally classified instances only and is of the form $< NP, VP, NP, min(score_i) >$.

2.3.3. Triple semantic analyzer

The resulting set of quads, thus pertains to our classified positive class (causal) instances. While it may be possible to judge the classified instances, by extending the test data annotation of the sentence to the causal phrase, it is better to validate the classified instances from the expert. In order to support the expert, with maximum information about the classified instances (since conversion from corpus to sentence and then to candidate phrases removes a large part of their context), we extend each NP in the classified quad, with its associated Concept Unique Identifier (CUI) and semantic type using the UMLS REST API.[1] This allows the system to identify if at-least one of the participating terms is semantically related to any medical terminology. If both terms do not have any corresponding concepts in UMLS, then it is also filtered out. The generation of this syntactically and semantically expanded set of classified instances then completes the process of lexical analysis and classification of the unseen clinical text.

2.4. Feedback loop

A feedback loop allows the expert to validate the classified instances produced by the MD module by using the semantic information expanding the noun phrases of the causal quads and the similarity score. The expert can indicate a phrase as causal or non-causal, providing a basis for updating the CTTM. This model evolution is acheived by generating the embedding vector for each of the expert validated causal classified instances, using BERT pre-trained models and simply appending the same to the training embedding vector list. Additionally, the phrases marked as non-causal, are added to a causal blocklist, which is then converted into an embedded vector, and compared with the vector lists of the CTTM. For each training embedding vector in the CTTM, if the similarity threshold with the blocklist embedded vector is greater than α_i, it is removed from the list. Initially, this lookup table is kept empty and as the expert identifies the correctness of causal phrases, it grows to include the correct phrases and discards similar non-causal phrases, for each of the six models. In this way, the CTTM evolves with each iteration and improves upon the previous results using expert feedback.

[1] https://documentation.uts.nlm.nih.gov/rest/home.html.

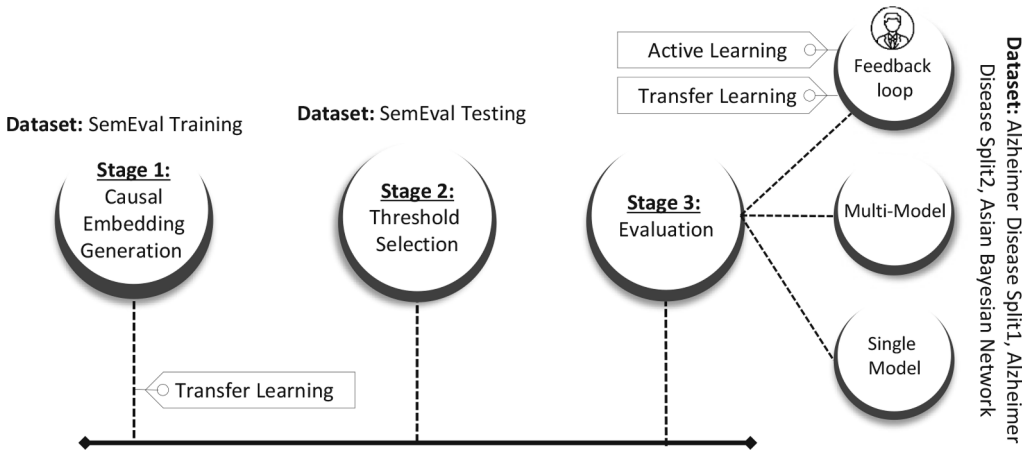

Fig. 4. Experimental Setup.

We validated the soundness of the proposed methodology by applying it on various datasets, and also compared the results with existing studies. As mentioned earlier, previous studies on causality classification have mainly focused on the creation and utilization of expert-generated rules. However, in a recent study [30], the authors presented a methodology, driven by the similarity between word embeddings, to classify instances from the same datasets we have used for evaluations. Their methodology is based on the identification of causal verbs between two labeled entities, followed by conversion of these verbs and those within the test data set into embedding vectors using Word2Vec. The embedding vectors are then compared using cosine similarity. If the similarity between the two vectors is greater than 0.5, the authors classify the verb from test instances as causal and add these verbs into the set of causal verbs used for subsequent matches. Finally using expert's rules, the authors classify the instance into one of the four different causality relationships (subject causes object, subject is the result of object, attribute relation, and certain relationship). The results presented by the authors indicate good performance of their model in comparison to two previous studies [23,25]. While the results presented by the authors in their manuscript are interesting, in their original form, they are incomparable to our results. More details of this gap in the evaluation strategies between the authors in [30] and our methodology is presented in Appendix D. We therefore, created an implementation of the Ning's strategy [30] to classify causal triples as causal or non-causal and compared the same with our results. During this implementation, we have utilized the same seed verb list, as presented by the authors in their research work, maintained the same similarity threshold value of 0.5, and followed the same design to classify each verb as causal. For any triple, where the verb was classified as causal, the triple is also considered as causal. The results comparison is shown in Section 4.3.

3. Experimental setup

The methodology presented in Section 2, represents a theoretical framework for identifying causal relationships in unstructured text. In order to build a sound realization of this framework, it is pertinent to identify the concrete models and algorithms, which can locally optimize each component, providing intermediate results with high performance and in turn amalgamate the workflows, providing a global optimal result for causality mining. Through various experiments we evaluated the impact of causal term expansion models, embedded vector generation methodologies, and similarity thresholds calculation to identify a well-balanced ecosystem, fulfilling our local and global optimization objectives.

Some initial experiments, including evaluation of only verb expansion, and embedding vector generation using Word2Vec, comparison of six pre-trained BERT models (base-nli-mean-tokens, large-nli-mean-tokens, base-nli-max-tokens, large-nli-max-tokens, base-nli-cls-tokens, large-nli-cls-tokens), and application of BioBert embeddings [42] are explained with some detail, in the Appendices A, B, and C, respectively. The rest of the experimental setup can be categorized into 3 stages, as shown in Fig. 4, where each following stage, receives data from all previous stages.

3.1. Stage 1 - Causal embedding generation

In Stage 1, Causal Embeddings were generated for the SemEval 2010 task 8 training dataset[32], using the six pre-trained BERT models. This dataset pertains to the semantic relation identification process and identifies the relationships between nominals for drug-drug interactions from biomedical texts. Each sentence in this training dataset and its counter part SemEval 2010 task 8 test dataset [32], is tagged with its most plausible truth-conditional interpretation using one of product-producer, content-container, cause-effect, and other semantic relations. However, since the target of this study is causality mining, we therefore, only considered the cause-effect tag as casual relation and all other as non-causal relations. The SemEval 2010 task 8 training dataset [32] comprises of 1003 causal sentences out of 8000 sentences. From these 1003 causal sentences, we extracted 1071 unique causal triples. The verb within each triple is then expanded using the pre-trained Google News model [33]. After the expansion, we take Cartesian product of the two encapsulating nouns of the source triple and one of the expanded verb to produce a little over 1.2 million expanded triples. Thus, with this expansion we are able to classify a wider range of causal relations, than what would have been possible, otherwise. Next we convert these expanded triples into embedding vectors using six pre-trained BERT NLI models [43,41], which include nli-base-mean-tokens, nli-large-mean-tokens, nli-base-max-tokens, nli-large-max-tokens, nli-base-cls-token, and nli-large-cls-token. These model differ in terms of their size (base or large) and the pooling layer used at the end of their deep neural network (mean pooling word tokens, max pooling word tokens, or cls pooling sentence token). Embedding vector generation for the 1.2 million expanded triples is a computationally expensive operation, which can take several days running on the CPU, however, due to the ability of the sentence_transformer library in python, to optimally use GPU, if available, the computational time is reduced, substantially. Through our experiments, we were able to process the expanded triples and produce the embedding vectors for base models in under 20 min each and for large models in an hour, each. Overall, the embedding vectors were produced in 4 h, using NVIDIA GeForce RTX 2060 GPU.

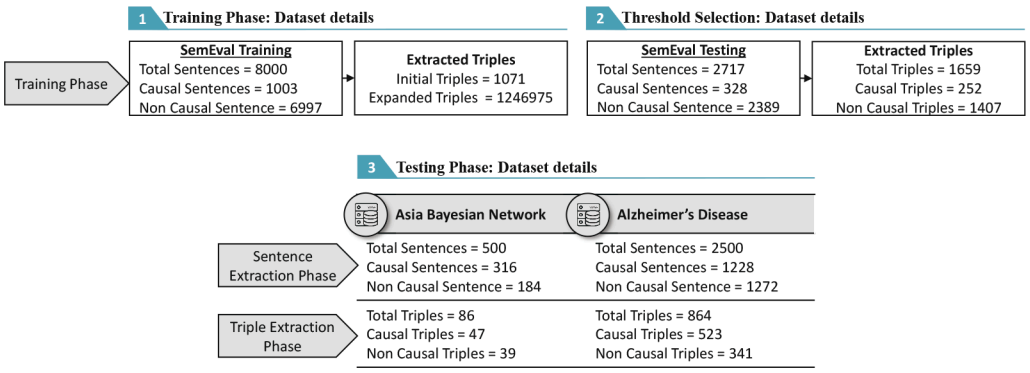

Fig. 5. Details of dataset.

3.2. Stage 2 - Threshold selection

Stage 2 is designed for threshold selection, whereby a sentence can be categorized as causal or non-causal, based on its similarity with the expanded triple set. Similarity threshold plays a vital role in the causality classification process and therefore requires extensive experimentation to select the best similarity score, above which a triple can be classified as causal. In order to fulfill this aim, we utilized SemEval 2010 Task 8 test dataset to learn the best threshold value, where the precision-recall curve (PRC) obtains maximum area under the curve. In biased datasets, where the ratio of positive class is much lower than the negative class, Area under the PRC (AUPRC) is an optimal metric for selecting the threshold [44]. As shown in Fig. 5, the SemEval 2010 Task 8 test dataset [32], contains 328 causal sentences, out of a 2717 total sentences (12.07% of positive class). We utilized AUPRC to learn optimal threshold values for each BERT models. The detailed result of threshold selection will be presented in Section 4.1.

3.3. Stage 3 - Evaluation

In Stage 3, we performed single model, multi-model, and multi-model with feedback loop evaluations on the Asian Bayesian Network dataset [30] and the risk factors of Alzheimer's disease (AD) [30], using the causal embedding vectors from Stage 1 and threshold values from Stage 2. The AD dataset consists of 1228 causal sentences out of 2500 sentences, while the Asian Bayesian Network dataset have 316 causal sentences from a set of 500 sentences. The sentences in these two datasets are tagged with either NP→NP (Noun Phrase influences Noun Phrase), NP-NP (Noun Phrase is related to Noun Phrase), or NP × NP (both nouns are irrelevant) label. In this study, we considered the first two tags (NP →NP and NP-NP) as causal and the remaining (NP × NP) as non-causal. Due to the large size of AD dataset and to test various iterations of the feeback loop, we split this dataset into two parts, using random selection for 50% partitioning. The complete AD dataset, contains 864 candidate triples, out of which 523 are causal (60.53%) and 332 are non-causal (39.47%). With 50% random split, the AD1 and AD2 dataset contain 432 triples each. AD1 contains 267 actual causal triples (61.80%) and AD2 contains 256 actual causal triples (59.26%). Evaluations by all three methodologies (single model, multi-model, and multi-model with feedback loop) were performed on these three instances of the datasets (AD1, AD2, and Asian Bayesian Network). This data split is especially, important to execute and evaluate multiple iterations of the feedback loop, on unseen data.

In single model, we evaluate the performance of each BERT model to check the effect of the model size in terms of base and large, and pooling strategies using CLS-token, mean of all output vectors, and max-over-time of the output vectors and select a single best performing model for causality mining. However, by inspecting the result of each BERT model in terms of unique causal triple identification via a very handy UpSet tool [45], which can plot associations between different sets and can be used to visualize relationships, where the traditional Venn diagrams may fail (such as when the number of sets are greater than 4).[2] Since the aim of our approach is to improve the accuracy of causal classification, even in presence of false positives, it is then pertinent to analyze the UpSet results, based on a "minimum" intersection degree metric. This entails, the evaluation of causal classifications for a minimum intersection degree such as degree ⩾1, degree ⩾2 and above. Intuitively, it can be seen that the performance results for degree ⩾2 should be less than the performance for degree ⩾1 and leads to a multi-model evaluation. The UpSet analysis performed in Section 4.2 revealed to used multi-model evaluation to increase efficiently of the causality mining.

In multi-model evaluations, we performed the experiments on the same three test datasets. However, in this case, we considered a triple as causal if any of the six BERT models tagged it as causal and non-causal otherwise. The results achieved in multi-model evaluation is shown in subSection 4.2.2.

Finally, we incorporated human expert's feedback into the multi-model similarity matching process, to analyze the change in the quality of causality detection. For this process, an expert (physician) from our collaborative hospital, verified the accuracy of the classified sentences. Since our automated process is dependent upon various datasets and has been repurposed, as explained earlier, this secondary verification is of utmost importance. This process was repeated in three iterations, while we ensured that once the CTTM is updated by the embedding vector of an expert verified causal triple, the same is not made a part of any subsequent test sets. Thus, the test sets in each iteration remain unseen. In Iteration-1, we used the embedded models (CTTM) trained on the SemEval 2010 Task 8 training dataset, and tested using the AD1 dataset. Embedding vectors corresponding to the correctly classified and expert verified causal and non-causal triples were then used to update the CTTM. In Iteration-2, this updated CTTM was then used to test the candidate triples from AD2 dataset. Once again, the correctly classified and expert verified causal and non-causal triples were used to again update the CTTM. Finally in iteration 3, the most recently updated version of the CTTM was then used for classifying the candidate triples from the Asia dataset. The details of the results achieved in each iteration are described in Section 4.2.3.

For experimentation, we used python code on Google Colab, with many additional libraries including Gensim models, NLTK, BERT sentence_tranformer, and sklearn. Using the same settings we developed a python based end-to-end application, which can extract causal relationships from an unseen copora. The application from Fig. 1 and its evaluation was run on a dedicated workstation with Intel(R) Core(TM)

[2] The interactive UI is available at http://vcg.github.io/upset/?dataset=10, with the data drescription file for our presented approaches present at https://raw.githubusercontent.com/Musarratpcr/CausalityDetection/master/Revision1/ADandAsianDatasetUpsetDescription.json.

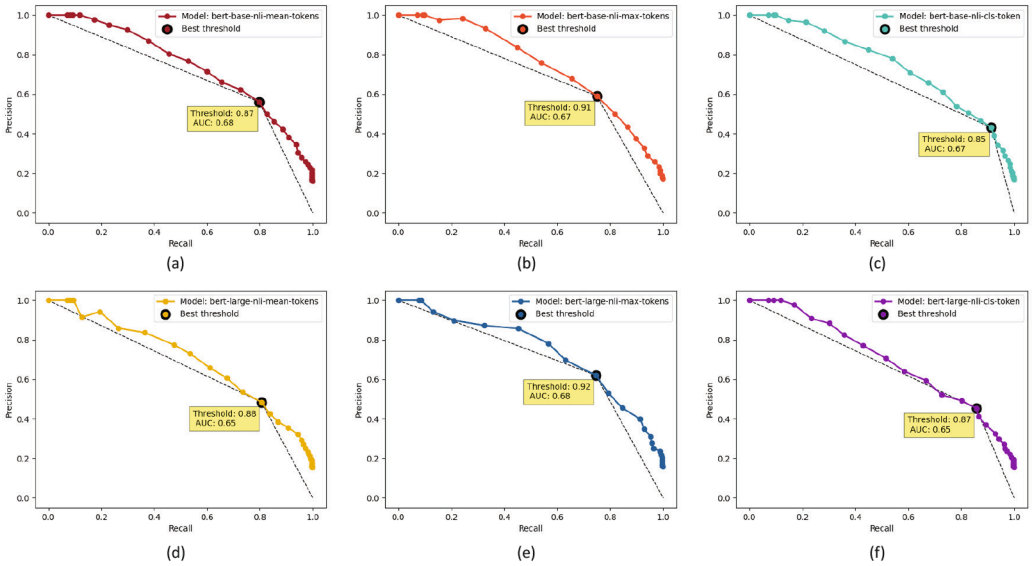

Fig. 6. Precision recall cure for threshold selection (a) bert-base-nli-mean-tokens (b) bert-base-nli-max-tokens (c) bert-base-nli-cls-tokens (d) bert-lart-nli-mean-tokens (e) bert-large-nli-max-tokens (f) bert-large-nli-cls-tokens.

Table 1
Application of trained embedding on Asia Bayesian Network dataset Legend: TP is True positive, FN is False Negative, FP is False Positive, TN is True Negative, A is accuracy, P is precision, R is recall, and F1 is F1 Score.

Scenario	TP	FN	FP	TN	A (%)	P (%)	R (%)	F1 (%)
BERT nli-base-mean-tokens	8	39	5	34	48.84	61.54	17.02	26.67
BERT nli-large-mean-tokens	37	10	26	13	58.14	58.73	78.72	67.27
BERT nli-base-max-tokens	9	38	10	29	44.19	47.37	19.14	27.27
BERT nli-large-max-tokens	21	26	14	25	53.49	60.00	44.68	51.22
BERT nli-base-cls-token	34	13	19	20	**62.79**	**64.15**	72.34	68.00
BERT nli-large-cls-token	38	9	26	13	59.30	59.38	**80.85**	**68.47**

Table 2
Application of trained embedding on Risk Factors of Alzheimer's Disease Split 1.

Scenario	TP	FN	FP	TN	A (%)	P (%)	R (%)	F1 (%)
BERT nli-base-mean-tokens	62	205	36	129	44.21	**63.27**	23.22	33.97
BERT nli-large-mean-tokens	111	156	80	85	45.37	58.12	41.57	48.47
BERT nli-base-max-tokens	72	195	45	120	44.44	61.54	26.97	37.50
BERT nli-large-max-tokens	80	187	54	111	44.21	59.70	29.96	39.90
BERT nli-base-cls-token	157	110	100	65	51.39	61.09	58.80	59.92
BERT nli-large-cls-token	165	102	104	61	**52.31**	61.34	**61.80**	**61.57**

Table 3
Application of trained embedding on Risk Factors of Alzheimer's Disease Split 2.

Scenario	TP	FN	FP	TN	A (%)	P (%)	R (%)	F1 (%)
BERT nli-base-mean-tokens	60	196	27	149	48.38	**68.97**	23.44	34.99
BERT nli-large-mean-tokens	128	128	70	106	54.17	64.65	50.00	56.39
BERT nli-base-max-tokens	74	182	37	139	49.31	66.67	28.91	40.33
BERT nli-large-max-tokens	88	168	54	122	48.61	61.97	34.38	44.22
BERT nli-base-cls-token	166	190	94	82	**57.41**	63.85	64.84	**64.34**
BERT nli-large-cls-token	172	84	111	65	54.86	60.78	**67.19**	63.82

i9-9900KF CPU, with 64 GB ram, and NVIDIA GeForce RTX 2060 GPU. The training model was produced in under 4 h, using a combination of CPU(for gensim based models which cannot use GPUs and are required for word expansion) and GPU(for BERT inference).

All code and results are available at the following link. https://github.com/Musarratpcr/CausalityDetection.

4. Results

In the following sub-sections, we shall provide the results obtained from various experiments in Stage 2 (threshold selection) and 3 (evaluation) of the setup(as shown in Fig. 4).

4.1. Stage 2 - Threshold selection

Following the process of preprocessing in Section 2.1, candidate triple extraction in Section 2.3.1, and causal candidate classification in Section 2.3.2, we calculated the cosine distance between the candidate triples of the SemEval 2010 Task 8 test dataset against the six BERT

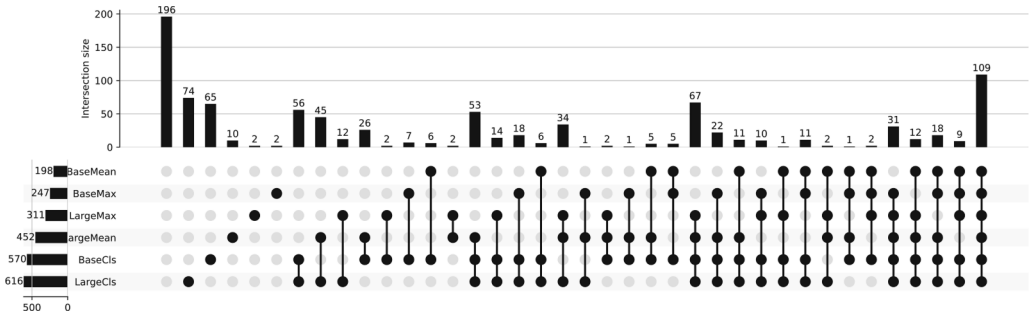

Fig. 7. UpSet analysis of BERT model classification coverage for a combined list of Risk Factors of Alzheimer's Disease and Asia Bayesian Network dataset.

Table 4
Application of Multimodel Embedding on Test Datasets.

Dataset	TP	FN	FP	TN	A (%)	P (%)	R (%)	F1 (%)
AD1	210	57	132	33	56.25	61.40	78.65	68.97
AD2	205	51	138	38	56.25	59.78	80.08	68.45
Asia	41	6	28	11	60.47	59.42	87.23	70.69

models. Then using the truth values (labels) of each candidate triple from the test dataset, and the similarity score pertaining to the cosine distance, we individually evaluated the six BERT models, producing charts shown in Fig. 6. We evaluated each threshold point, by connecting it with the inverse diagonal of the graph (From Precision = 1 and recall = 0 to Precision = 0 and recall = 1). We then calculated the area under this newly formed curve, and found out the threshold where this area was maximized. The average threshold value α then comes to 0.88, however, utilizing this average value in the multi-model CTTM would greatly affect the performance, by misclassifying instances for five individual models (more phrases will be classified as causal by bert-base-nli-mean-tokens, bert-base-nli-cls-token, bert-large-nli-cls-token, and less for bert-base-nli-max-tokens, and bert-large-nli-max-tokens). Instead, in the CTTM, we utilized the individual threshold values of each BERT model α_i, to classify instances, when compared to the corresponding embedding vector list.

4.2. Stage 3 - Evaluation

In order to evaluate the performance of our generated triples and the selected threshold we then performed single-model, multi-model, and multi-model with feedback loop evaluation of three, as yet, unseen datasets, the Asian Bayesian Network dataset and the two partitions for risk factors of Alzheimer's disease (AD1 and AD2). These are discussed as follows:

4.2.1. Single model evaluation

In the Asia Bayesian Network dataset from a total of 86 qualifying triples, 47 are actual causal(54.65%) and 38 are non-causal(44.18%). The results achieved by each BERT model on this dataset are shown in Table 1. BERT models, utilizing the complete phrase as a token and then cls for pooling at the final layer, show good performance, when compared with the others. Overall, the best values for accuracy, precision, recall and F1 are achieved by these models, however, the BERT nli-large-mean-tokens closely follows the classification performance. However, the results for base models with mean tokens and max tokens, indicate very bad performance with F1 measure under 28% (caused by the low performance of recall obtained by these models).

In absolute terms, the classification performance for the AD dataset in terms of accuracy, and F1 measure is lower than the Asia Bayesian Network dataset, at par for precision, and higher for recall. Comparison amongst the six models shows some similarity with the previous results. Causal classification of AD1 shown in Table 2, achieves better performance, in terms of its accuracy, recall, and F1 for the two cls-token models, with the large version achieving the best results. The performance of other models, lacks behind substantially with F1 rates between 34% and 49%. The precision rates of these six models, are however, within 5.15 percentage points, which indicates that the ability of each model to correctly identify the actual causal phrase, when a triple is classified as causal, is similarly good (or bad). Another important metric to analyze these results is to look at the recall rates, which in the case of Asia Bayesian Network dataset, were able to correctly identify 80.85% of the actual causal instances, however, for AD1 only identify 61.80%, in the best case. For the AD2 dataset, performance metrics shown in Table 3, indicate the best recall rate of 67.19%, which is better than the results for AD1 but substantially smaller than Asia Bayesian Network

Table 5
Feedback loop results on test datasets.

Iteration	Dataset	Dataset Evaluation				Expert Evaluation	
		A	P	R	F1	Added to Embeddings	Added to Block List
1	AD1	56.25%	61.40%	78.65%	68.97%	314	28
2	AD2	60.88% (↑ 4.63)	60.43% (↑ 0.65)	98.44% (↑ 18.36)	74.89% (↑ 6.44)	268	49
3	Asia	61.63% (↑ 1.16)	60.00% (↑ 0.58)	89.36% (↑ 2.13)	71.79% (↑ 1.1)	58	12

A: Accuracy, P: Precision, R: Recall, The values in parenthesis represent rate of change from multi-model results.

Table 6
Result comparison with Ning's method on test datasets.

Dataset	Ning's Method Evaluation				Proposed Method Evaluation			
	A(%)	P(%)	R(%)	F1(%)	A(%)	P(%)	R(%)	F1(%)
AD1	61.81	61.81	100	76.39	56.25	61.40	78.65	68.97
AD2	59.26	59.26	100	47.42	60.88	60.43	98.44	74.89
Asia	54.65	54.65	100	70.68	61.63	60.00	89.36	71.79

dataset. The best F1 rates for AD2 are achieved by the base version of the BERT cls token based model.

These results provide empirical proof for causality detection, based on causal phrase extraction and expansion. However, selection of a single model, based on these results alone, would not resolve the problem of causality detection in clinical text, where it is critical to identify most if not all actual causal sentences. Hence a deeper look at the coverage of these six BERT models, in terms of correctly classifying the actual causal instances is necessary.

The associations between the results achieved by six bert models on combined triple phrases from Asia Bayesian Network and Risk Factors of Alzheimer's Disease datasets is shown in Fig. 7. Amongst the 950 candidate triples, 754 have been classified as causal by one or more of the BERT NLI models. The Base-Mean classifier, is unable to uniquely classify any candidate triple as causal, however, the other five models, classify 153 instances as causal. With classification interSection 2, 156 candidate triples are classified by a combination of only two models uniquely classify an instance as causal. Extending this calculation on the numbers achieved via UpSet analysis, unique coverage rate from degree 1–6 are 153(20.29%), 156(20.69%), 139 (18.44%), 127 (16.84%), 70 (9.28%), and 109 (14.46%), respectively. The actual causal triples in the 950 candidate triples are 570. True positive classification numbers for the six models with degree 1–6 are 83 (14.56%), 96 (16.84%), 86 (15.09%), 76 (13.33%), 45 (7.89%), and 70 (12.28%), respectively. These results indicate that single model application of anyone of the six BERT models, will have low candidate classification coverage and even lower true positive rates.

4.2.2. Multi-model evaluation

Theoretical analysis of the results shown in Fig. 7, indicate that when degree ⩾6, 109 phrases have been classified as causal, out of which 70 are actual causal. The accuracy of this classification is 43.26% and F-1 rate is 20.62%. For degree ⩾5, 179 phrases have been classified as causal, with 115 as true positive. The accuracy rate now, increases to 45.37%, while the F-1 goes up to 30.71%. Similarly, for degree ⩾4 accuracy further increases to 48.00% and F-1 to 43.61%. For degree ⩾3, the accuracy becomes 51.47%, and F-1 54.58%. For degree ⩾2, accuracy further improves to 55.26%, and F-1 to 63.71%. Finally for degree ⩾1, at least 1 model classified 754 instances as causal, out of which 456 are actual causal. The accuracy increases to 56.63%, and F-1 to 68.88%. Matching the intuition, presented earlier, this analysis, also shows, that if at least one model classifies a candidate phrase as causal, it should be accepted, to achieve the highest realistic performance.

Practical application of the multi-model methodology, where a phrase is considered causal, if at least one model classifies it as such, produces the same result, showing an accuracy rate of 56.63% and F-1 as 68.88% for the combined dataset. Separately, the results for Asia Bayesian Net dataset show small improvement in their F-1 score (multi-model selected additional 3 correct causal phrases than the best results for BERT nli-large-cls-token on this dataset) and a slight drop in its accuracy, due to an increases number of True negatives (skewing the accuracy measure, towards positive results). Both AD1 and AD2 dataset, show substantial improvement of causality classification, with the application of multi-model technique. The number of correctly classified causal pharses in AD 1 have increased form 165 in the best case to 210, while for the AD2 have increase from 172 in the best single model application to 205 here. Overall the performance of multi-model classification on this dataset has brought it at par with the result of the other dataset. The F-1 measures for both AD1 and AD2 have increased, showing the correctness of the multi-model strategy for causality detection. However, the large number of false positives and true negatives, still leave a room for improvement of this model, which we resolved by additionally employing the feedback loop. The results for this upgrade are shared in the following evaluation.

4.2.3. The feedback loop evaluation

In Iteration-1, the CTTM trained on the SemEval 2010 Task 8 training dataset was tested using the AD1 dataset. The six models in CTTM were updated by adding embedded vectors for the 314 causal triples verified by the expert and, removal of triples with similarity score α_i for the 28 marked as incorrectly classified. In the base version of the nli-mean-tokens model, 60 similar triples were removed, while in large version 190 triples were removed. Similarly, for the base and large version of the nli-max-tokens 94 and 143 triples were removed,respectively. Finally for the cls-token version, 676 triples were removed from base and 626 triples from the large version. As shown in the Table 5, the accuracy of multi-model CTTM application on the AD2 dataset, shows minor improvement, in accuracy (from 56.25% to 60.87%), precision (from 59.78% to 60.43%), recall(80.08% to 98.44%), and F1 (68.45% to 74.86%), on incorporation of results from AD1.

In Iteration-2, the expert verified 368 classified causal triples as correct, while 49 were marked as non-causal. Based on this new set of causal triples, we again updated the CTTM before iteration 3, to further add the 368 embedding vectors and removed 175 triples from base version of the nli-mean-tokens, and 308 from the large version. For the nli-max-tokens 251 were removed from base version and 477 from large version. Finally, in the case of cls-token 804 were removed from base and 774 from large.

In Iteration-3, the evolved CTTM was applied on the Asia Bayesian Network Dataset, which registered small improvements on the multi-model results. Since this dataset is the smallest of the three, CTTM model evolution has very little impact on it. Addition of 759 triples in the original 1,246,975 embedded vectors from CTTM model before iteration 1, and removal of various others (between the minimum total of 235 triples removal from base nli-mean-tokens in 2 iterations and maximum of 1480 from base cls-token), increased the true positive from 38 in best case single model to 41 in multi-model, and finally to 42 in the third iteration.

4.3. Comparison with existing studies

In order to compare our methodology with an existing study, we utilized the methodology presented by [30] to classify sentences as causal or non-causal, from the AD1, AD2, and Asia dataset. However, since our methodology incorporates the datasets into the CTTM, using feedback loop, and because we want to maintain the unseen nature of these, so as not to contaminate the results, we compared our iteration 1 result for AD1, iteration 2 result for AD2, and iteration 3 result for Asia dataset. At these specific points, the datasets are unseen and true test sets. The results for causal classification on the test datasets for both methodologies (Ning's and proposed) are shown in Table 6. We observed that our implementation of Ning's methodology [30], classifies all triples as causal achieving a recall rate of 100%. However, the accuracy, precision, and F_1 scores are decreasing by comparatively large margins. These results are in line with our previous experiments based on word embedding (Appendix A). Hence, it is safe to conclude that even when starting with a well-identified set of causal verbs, word embedding by itself is not sufficiently able to evolve the causality classification model. On the other hand, our methodology is able to improve upon its results across iterations.

5. Discussion

The main aim of this study is to develop a framework that can identify causal sentences in clinical text. The success criteria of this framework are dependent on correctly identifying most causal relationships, with some leeway available in incorrect classification of non-causal sentences as causal. Precision, recall, and their association in the form of F1 provides the metric to evaluate our proposed framework, in parts, as a whole, and with existing work. Application of this classification methodology can then enable an expert from the domain of

healthcare and wellness, to be able to contextually summarize the contents of the clinical text. To this end, we extract the causal phrases from the causal sentences, which are larger in numbers but smaller in their participating linguistic elements (including two NP and one VP).

The results presented in Section 4, provide the performance metrics for various steps leading up to our proposed multi-model classification with a feedback loop. In particular, the evaluation metrics for the Asian Bayesian Network, and two partitions of the Alzheimer's disease datasets, generally saw an increase, when moving from single model to multi-model and then to multi-model with a feedback loop.

The rationale for moving from using a single BERT NLI model to a multi-model application was established using UpSet analysis, presented in Section 4.2.1. Additionally, the rationale for moving from multi-model to multi-model with a feedback loop can be naively established from intuition, however, it is far more beneficial to analyze the phrases which were originally classified by the machine learning models and then removed by the expert. As an example one of the triples identified by the multi-modal methodology from the AD1 dataset is "cancer = alcohol". The origin of this triple can be traced back to the following instance:

"After adjusting for various socioeconomic and health variables, no significant differences were observed between hazardous drinking and type of cancer [PR = 0.99 = 0.83–1.17) in people with alcohol-related cancers compared to non-alcohol related cancers] and time since diagnosis [PR = 1.01 in people with a cancer diagnosed >5 years ago compared to those diagnosed <=5 years ago]."

Stanford POS tagger (version 3.9.2) had incorrectly identified the symbol "=" as a feasible VP and since this fell between the two entities (cancer and alcohol), this triple was considered valid. The expert verified this triple as incorrect since it does not provide enough information to classify the original sentence as causal or non-causal. Hence the embedding for this triple and all others similar to it, with the threshold equal to or above, for each model, were then removed. This removal process is not dependent only on the VP, as in iteration 2, we observed additional triples with invalid VPs, such as "stroke = diabetes" and "alcohol [depression". In iteration 3, none of the triples had a symbol as a VP.

Another triple identified by the expert as incorrect was "smoke monitored hypertension". This triple contains a valid VP, tagged by the POS tagger as "VBN". The original instance from which this triple was extracted is as follows.

"The earlier advice to physicians still seems prudent and is briefly stated: 1) Try to avoid prescribing oral contraceptives for women over 35 years of age; 2) Women who smoke cigarettes should avoid using oral contraceptives, and users should not smoke; 3) Prescribe the formulation with the lowest dose and/or potency of estrogen that is effective and that does not cause unacceptable "breakthrough" bleeding; 4) Women with hypertension should be carefully monitored, and women who develop hypertension while on oral contraceptives should be switched to another form of contraception, if possible."

In hindsight, intuitively, it is evident from the original text that the phrases "smoke", "monitored", and "hypertension" all belong to different contexts. However, the machine learning models are agnostic to such contexts, unless they can incorporate a very large number of sentence and document structuring rules. While there are other triples extracted from this instance, which may qualify the instance as causal or non-causal, it does not help fulfill our aim of identifying individual causal sentences from the classification of causal triples. Hence, we update the model, to only hold those triples which can represent causal relationships, from a wide variety of datasets. In iteration 3, only 12 triples were identified as incorrect by the expert, including "bronchitis smoke smoking" and "lung cancer secondhand smoking" (while lung cancer can be caused by secondhand smoking, this triple is missing the causal verb). Here it is pertinent to mention that by removing these triples, we are not changing our results but rather evolving the model for subsequent classification in unseen datasets. In the absence of active learning, our model would not be able to update itself and hence provide relatively mediocre results as discussed in Section 4.2.2.

In iterations 2 and 3, incorrect triple embeddings similar to the ones identified in previous iterations are not included. The similarity is determined by converting the incorrect triples into embedding vectors and using the 6 BERT models to determine all embeddings which have cosine similarity above their respective thresholds. The correctly identified causal triples are added into the CTTM by appending their embeddings at the end. Additionally, on subsequent classifications, the data instances (a sentence, text excerpt, or a document) are classified using the evolved CTTM. This is why even after removing related embeddings the results obtained by including active learning are gradually increasing, even on unseen and minimally related datasets (AD2 in iteration 2 to Asia in iteration 3).

On a related note, the evaluation of our results has been performed using the labels of the test data, while the expert-provided feedback was used only to update the model. As a result some phrases such as "cancer associated alcohol", and "cancer rising alcohol", were classified by the machine learning model and the expert as causal, however, the dataset had the associated sentence labeled as non-causal. Since CTTM is direction agnostic it is unable to distinguish between various forms of the causal phrases such as "cause triggers effect" and "effect triggered by cause". As an example the triple "cancer associated alcohol" has been extracted from the following instance:

"The results showed that frequent intake of fruits, chicken, fish and alcohol drinking were associated with risk for colorectal cancer."

Here, the triple has been correctly identified, since one of the verbs between "cancer" and "alcohol" is "associated". The triple was also identified as causal by the CTTM and the expert, however, the dataset marks it as non-causal. Thus while the triple itself is causal, the originating instance is non-causal (the dataset labels it as "cancer x alcohol"), which negatively affects the evaluations and reduces our performance.

The causality classification methodology presented in this manuscript attempts to alleviate problems caused by discrepancies in causally valid POS tagging, triple expansion (which can include non-causal triples), and other operations. Through the use of active learning, we have observed an increase in the performance of our proposed methodology. While, we have reduced the expert's involvement in the causality classification process, substantially, when compared with the previous studies, further reduction is possible through the use of specialized POS taggers, contextual triple expansions, better sentence embedding generation, and similarity measures.

6. Conclusion

Active transfer learning using amalgamation of results from multiple models is a novel and, as proved above, successful methodology for identifying causal sentences. This two class classification problem, whereby we aimed to correctly identify the causal sentences, shows better and maintainable recall rates, while the performance of this methodology, in terms of accuracy, precision, and F measure can be improved by incorporating additional active learning iterations. In future, we shall look towards the application of our methodology for solving other relevant clinical problems, such as generation of patient summaries from clinical text.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center)

support program (IITP-2017-0-01629) supervised by the IITP (Institute for Information & communications Technology Promotion), by the Korea government MSIT (Ministry of Science and ICT) grant (No. 2017-0-00655), by the MSIT Korea, under the Grant Information Technology Research Center support program (IITP-2020-0-01489), (IITP-2021-0-00979) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation) and NRF2019R1A2C2090504.

Appendix A. Experimental result with Word2Vec embeddings

In Experiment Appendix A, we performed some initial experiments to test the applicability and performance of Word2Vec based embedding vector generation process, for causal verbs and causal triples, in both training and test datasets. A summary of the results are shown in Table A.7.

In Experiment 1, we extracted the causal verbs using the stanford POS tagger, from our training dataset. Without any expansion, we then applied word embedding on the causal verb, which was used to look up similar verbs in the SemEval test data set. In this iteration, we predicted 1318 sentences to be positively causal and 1399 sentences to be non-causal. From the predicted positive sentences, actual causal sentences were 205, and incorrect ones were 1113. The accuracy of this approach is 54.50% and recall 62.5%. However, the precision of this scenario is only 15.55% and F1 is 24.81%.

In Experiment 2, we expanded the causal verbs extracted in experiment 1 using Google News pre-trained model. Using word embedding, we transformed the extracted as well as the expanded causal verbs into word vectors. In the SemEval test data, using cosine similarity, 1453 sentences were classified as causal, with 210 correctly classified and 1243 incorrectly. After causal verb expansion the accuracy was dropped to 49.90%, precision to 14.45%, F1 to 23.58% but recall increased slightly to 64.40%. This indicates that word expansion from Google News pre-trained model has a very small impact on the classification process.

In Experiment 3, we switched the word expansion model to ConceptNet, with numberbatch embeddings, which provides semantically similar terms. In this iteration, we predicted 929 sentences to be causal and 1788 sentences to be non-causal. However, only 59 causal sentences were correctly predicted, with an accuracy of 58.07%, recall of 17.98% and lowest precision of 6.35% and F1 of 09.39% amongst all experiments. Causal terms are highly discriminable, while the words expanded with ConceptNet have higher diversity and lacks discrimination, which leads to the drastic decrease in the model performance [46]. The results obtained thus far have proved the in-applicability of Word2Vec based embedding vectors generation. The Word2Vec considered a word without its context and neighbor terms, which may lead to inappropriate vector generation. Therefore, we generated the embedding vectors via BERT models in the upcoming experiments.

Table A.7
Initial Experiments with Word2Vec based embedding vector generation on SemEval Test dataset.

Experiment	TP	FN	FP	TN	A (%)	P (%)	R (%)	F1 (%)
1	205	123	1113	1276	54.50	15.55	62.5	24.81
2	210	118	1243	1146	49.90	14.45	64.02	23.58
3	59	269	870	1519	58.07	06.35	17.98	09.39

Appendix B. Experimental result with BERT embeddings

In Experiment Appendix B, like in the experiment Appendix A only verb was expanded. However, is this experiment the embedding vectors were generated using 6 BERT models to utilize sentence level embedding vector generation for a more contextual comparison. We compared 6 different BERT pre-trained models in terms of their performance on our test data set, with summary results shown in Table B.8 [43,41]. The 6 BERT models (nli-base-mean-tokens, nli-large-mean-tokens, nli-base-max-tokens, nli-large-max-tokens, nli-base-cls-token, and nli-large-cls-token) differ in terms of their model size(base or large) and the pooling layer used at the end of their deep neural network(mean pooling word tokens, max pooling word tokens, or cls pooling sentence token). Experiment 4 pertains to the base form of the BERT model that uses mean token pooling, while Experiment 5 uses the large form of similar layered model. Likewise, Experiment 6 is the base model, while Experiment 7 is the large model, with max pooling layer. Finally, Experiment 8, and 9 are base and large models, respectively, with cls pooling layer. The result obtains in each experiments is shown in Table B.8.

The result of these experiments show much improved performance, with experiment 4 (base model with mean pooling) showing the best accuracy (88.55%), precision(52.27%) and F1 (55.76%). The best recall(69.82%), is however, produced by the experiment 7 (large model with max pooling). On close inspection, we found experiment 7 to have correctly classified 229 sentences out of which 196 sentences were exactly similar to the True Positive results in experiment 4. However, the precision of experiment 7 is relatively small, due to the large number of False Positives.

Beyond these tests, it is also imperative that the generated embedding are tested on other text corpora for determining their ability to maintain acceptable performance, generally. Asia Bayesian Network and risk factors of Alzheimer's disease (AD) dataset were used to test this generalization. The results for the former are shown in Table B.9 and later in Table B.10. As shown in Table B.10 accuracy of each model decreases on Asia Bayesian Network as well as AD datasets. However, precision as well as recall of models shows a slight improvement on diverse datasets. In results for Experiment A.2 on the Asia Bayesian Network dataset, BERT nli-base-mean-tokens and BERT nli-large-mean-tokens show a precision of 100%, which is because of 0 false positives, however, this result is biased due to the very small number of identified causal triples.

These results paint an abysmal picture of the Experiment A.2 process. This is due to the fact that the verbs identified as causal through extraction from SemEval training dataset and their expansion are not able to capture all the causal sentences. These result partially support our novel methodology of incorporating the nominals (nouns and noun phrases) in the text producing the embedded vectors, thereby switching to causal quads for causal sentence identification. The intuition behind this arrangement, stems from the fact that causal sentences, implicitly contain semantic relationships between the cause and effect entities. Addition of these entities in the causal relationship identification process would spread a wider net for causal sentence identification. This intuition has been materialized and empirically tested in the manuscript.

Table B.8
Experiment A.2 - Setting 2 with BERT based embedding vector generation on SemEval Test dataset.

Experiment	TP	FN	FP	TN	A (%)	P (%)	R (%)	F1 (%)
4 - BERT-base-nli-mean-tokens	196	132	179	2210	**88.55**	**52.27**	59.76	**55.76**
5 - BERT-large-nli-mean-tokens	211	117	300	2089	84.65	41.29	64.33	50.30
6 - BERT-base-nli-max-tokens	227	101	633	1756	72.98	26.40	69.21	38.22
7 - BERT-large-nli-max-tokens	229	99	564	1825	75.60	28.88	**69.82**	40.86
8 - BERT-base-nli-cls-token	202	126	217	2172	87.38	48.21	61.59	54.08
9 - BERT-large-nli-cls-token	206	122	264	2125	85.79	43.83	62.80	51.63

Table B.9
Experiment A.2 - Application of trained embedding on Asia Bayesian Network dataset.

Scenario	TP	FN	FP	TN	A (%)	P (%)	R (%)	F1 (%)
BERT nli-base-mean-tokens	2	45	0	38	47.06	100.00	4.26	08.16
BERT nli-large-mean-tokens	4	43	0	38	49.41	100.00	8.51	15.69
BERT nli-base-max-tokens	11	36	11	27	44.71	50.00	23.40	31.88
BERT nli-large-max-tokens	31	16	18	20	**60.00**	63.27	**65.96**	**64.58**
BERT nli-base-cls-token	6	41	1	37	50.59	**85.71**	12.77	22.22
BERT nli-large-cls-token	1	46	2	36	43.53	33.33	2.13	04.00

Table B.10
Experiment A.2 - Application of trained embedding on Risk Factors of Alzheimer's Disease dataset.

Scenario	TP	FN	FP	TN	A (%)	P (%)	R (%)	F1 (%)
BERT nli-base-mean-tokens	53	423	16	316	45.67	**76.81**	11.13	19.45
BERT nli-large-mean-tokens	162	314	83	249	50.87	66.12	34.03	44.94
BERT nli-base-max-tokens	276	200	148	184	**56.93**	65.09	57.98	**61.33**
BERT nli-large-max-tokens	282	194	194	138	51.98	59.24	**59.24**	59.24
BERT nli-base-cls-token	110	366	50	282	48.51	68.75	23.11	34.59
BERT nli-large-cls-token	176	300	84	248	52.48	67.69	36.97	47.83

Appendix C. Experimental result with BioBERT embeddings

The experiments performed in Appendix B are repeated by replace the BERT model with BioBert for generated trigger and candidate embeddings for comparing their similarities. As mentioned earlier, the trigger in the form of triple < noun, verb, noun > was extracted from SemEval training datasets, and the verb terms were expanded with Google news model to extended the converge of the triggers. We calculate precision recall curve as shown in Fig. C.8, to identify the similarity cut off value of 0.96 for classifying a triple as causal and non-casual. However, the performance of the BioBert Embeddings are very low on the test dataset as shown in Table C.11. The unexpected performance of the BioBERT based embeddings is mainly due to the fact our test dataset contains non-clinical concepts along with clinical concepts. Therefore, we used Bert models instead of BioBert for our experiments and evaluations.

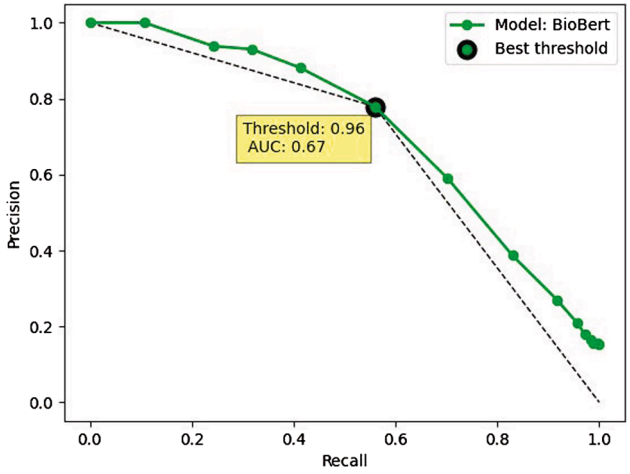

Fig. C.8. Precision recall curve for threshold selection for BioBert.

Table C.11
Application of BioBert Embedding on Test Datasets.

Dataset	TP	FN	FP	TN	A (%)	P (%)	R (%)	F1 (%)
Asia	5	42	6	33	44.19	45.45	10.64	17.24
AD	29	494	10	331	41.67	74.36	05.54	10.32

Appendix D. Comparison with existing studies

Using the same datasets (Asian Bayesian Network and risk factors of Alzheimer's disease), previous studies have utilized rule based and word embedding based strategies to classify the participating instances (each instance contains one or more sentences) into causal or non-causal sets. The results in terms of their precision, recall, accuracy, and F_1-score is shown in Table D.12. These results indicate that the three studies ([23,25,30]) are able to accurately classify between 74% and 76% of instances in the Asia dataset and between 75% and 83% of the AD dataset. In all three of these methodologies, the researchers have utilized the labeled entities to classify the instance as causal, if they follow rules generated by experts or contain a causal verb. Contrarily, our methodology extracts and classifies the triples with each sentence of the dataset, as shown in Table 4. While in the previous studies the two entities can be classified as causal based on their existential semantics in the corresponding instance and non-causal otherwise, our methodology performs independent classification. Summarily, while the previous studies perform causal classification within the context of the instance, our methodology provides generic classification, within a limited context. Resultantly, the results obtained by our methodology and the ones presented in Table D.12 are not comparable, unless a mapping from triples to the instances can be made.

The performance comparison between previous studies and our methodology for the Asia and AD datasets is not straight-forward. Firstly, the causality classification aims of the previous studies on these datasets and our presented methodology is not same. Secondly, since the feedback loop in our methodology incorporates the causal triples from the two datasets into the set of embedding vectors, it is important to perform the comparison on post-multi-model and pre-feedback-loop versions of our model. Thirdly, we discarded the set of triples from the test dataset in favor of a list of triples and their corresponding instances from the datasets. This extended data-structure allows the evaluation of four strategies to map the task of causal triple classification on to instance classification (identification of causal relationship between labeled entities within the scope of the instance). The four strategies include, causal instance on one causal triple (classify the instance as containing the causal relationship, if at least one triple is classified as causal by our methodology), causal instance on half causal triples (if at least half of the triples extracted from the instance are causal), causal instance on maximum causal triples (if more than half of the triples extracted from the instance are causal), and finally, causal instance on all causal triples. The results for these four strategies are shown in Table D.13.

Table D.12
Causality classification results of existing methods [30].

Method	Asia Network Dataset				AD Dataset			
	A(%)	P(%)	R(%)	F1(%)	A(%)	P(%)	R(%)	F1(%)
Bui's Method	74.60	70.66	78.21	74.24	75.50	73.43	76.11	74.75
Alashri's Method	74.20	67.83	79.19	73.07	78.32	73.84	81.92	77.67
Ning's Method	76.20	70.20	78.90	74.30	82.60	80.68	85.54	83.04

Table D.13
Instance classification result for Asian and AD datasets.

Strategy	Asia Network Dataset				AD Dataset			
	A(%)	P(%)	R(%)	F1(%)	A(%)	P(%)	R(%)	F1(%)
Causal instance on one causal triple	57.60	61.25	76.61	68.07	57.12	59.68	39.17	47.30
Causal instance on half causal triples	58.20	61.75	76.61	68.38	57.08	59.63	39.09	47.22
Causal instance on maximum causal triples	58.20	62.01	75.25	67.99	57.36	60.31	38.60	47.07
Causal instance on all causal triples	58.80	63.44	71.18	67.09	57.28	60.39	37.87	46.55

References

[1] C. Puente, A. Sobrino, J.A. Olivas, E. Garrido, Summarizing information by means of causal sentences through causal graphs, J. Appl. Logic 24 (2017) 3–14.

[2] P. Li, K. Mao, Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts, Expert Syst. Appl. 115 (2019) 512–523.

[3] N. Asghar, Automatic extraction of causal relations from natural language texts: a comprehensive survey, arXiv preprint arXiv: 1605.07895 (2016).

[4] L.J. Jensen, J. Saric, P. Bork, Literature mining for the biologist: from information retrieval to biological discovery, Nat. Rev. Genet. 7 (2) (2006) 119–129.

[5] R. Girju, Automatic detection of causal relations for question answering, in: Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering, vol. 12, Association for Computational Linguistics, 2003, pp. 76–83.

[6] C. Hashimoto, K. Torisawa, J. Kloetzer, M. Sano, I. Varga, J.-H. Oh, Y. Kidawara, Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, vol. 1: Long Papers, 2014, pp. 987–997.

[7] T.N. De Silva, X. Zhibo, Z. Rui, M. Kezhi, Causal relation identification using convolutional neural networks and knowledge based features, World Acad. Sci. Eng. Technol. Int. J. Comput. Electr. Autom. Control Inform. Eng. 11 (6) (2017) 696–701.

[8] S. Zhao, T. Liu, S. Zhao, Y. Chen, J.-Y. Nie, Event causality extraction based on connectives analysis, Neurocomputing 173 (2016) 1943–1950.

[9] S. Doan, E.W. Yang, S.S. Tilak, P.W. Li, D.S. Zisook, M. Torii, Extracting health-related causality from twitter messages using natural language processing, BMC Med. Informat. Decision Making 19 (3) (2019) 79.

[10] C.F. Meyer, Introducing English Linguistics International, student ed., Cambridge University Press, 2010.

[11] E.M. Ponti, H. O'horan, Y. Berzak, I. Vulić, R. Reichart, T. Poibeau, E. Shutova, A. Korhonen, Modeling language variation and universals: A survey on typological linguistics for natural language processing, Comput. Linguist. 45 (3) (2019) 559–601.

[12] L. Torrey, J. Shavlik, Transfer learning, in: Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, IGI Global, 2010, pp. 242–264.

[13] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2009) 1345–1359.

[14] K. Weiss, T.M. Khoshgoftaar, D. Wang, A survey of transfer learning, J. Big Data 3 (1) (2016) 9.
[15] F. Olsson, A literature survey of active machine learning in the context of natural language processing, 2009.
[16] B. Settles, From theories to queries: Active learning in practice, in: Active Learning and Experimental Design workshop In conjunction with AISTATS 2010, 2011, pp. 1–18.
[17] L. Zhao, S.J. Pan, E.W. Xiang, E. Zhong, Z. Lu, Q. Yang, Active transfer learning for cross-system recommendation, in: Twenty-Seventh AAAI Conference on Artificial Intelligence, 2013.
[18] X. Wang, T.-K. Huang, J. Schneider, Active transfer learning under model shift, in: International Conference on Machine Learning, 2014, pp. 1305–1313.
[19] R. Girju, D.I. Moldovan, et al., Text mining for causal relations, in: FLAIRS conference, 2002, pp. 360–364.
[20] R.M. Kaplan, G. Berry-Rogghe, Knowledge-based acquisition of causal relationships in text, Knowl. Acquisit. 3 (3) (1991) 317–337.
[21] K. Raja, S. Subramani, J. Natarajan, Ppinterfinder—a mining tool for extracting causal relations on human proteins from literature, Database 2013 (2013).
[22] G.A. Miller, Wordnet: A lexical database for english, Commun. ACM 38 (11) (1995) 39–41, https://doi.org/10.1145/219717.219748.
[23] Q.-C. Bui, B.Ó. Nualláin, C.A. Boucher, P.M. Sloot, Extracting causal relations on hiv drug resistance from literature, BMC Bioinformat. 11 (1) (2010) 1–11.
[24] S.V. Cole, M.D. Royal, M.G. Valtorta, M.N. Huhns, J.B. Bowles, A lightweight tool for automatically extracting causal relationships from text, in: Proceedings of the IEEE SoutheastCon 2006, IEEE, 2006, pp. 125–129.
[25] S. Alashri, J.-Y. Tsai, A.R. Koppela, H. Davulcu, Snowball: extracting causal chains from climate change text corpora, in: 2018 1st International Conference on Data Intelligence and Security (ICDIS), IEEE, 2018, pp. 234–241.
[26] D.-S. Chang, K.-S. Choi, Causal relation extraction using cue phrase and lexical pair probabilities, in: International Conference on Natural Language Processing, Springer, 2004, pp. 61–70.
[27] E. Blanco, N. Castell, D.I. Moldovan, Causal relation extraction, in: Lrec, 2008.
[28] D. Zeng, K. Liu, S. Lai, G. Zhou, J. Zhao, et al., Relation classification via convolutional deep neural network (2014).
[29] T.H. Nguyen, R. Grishman, Relation extraction: Perspective from convolutional neural networks, in: Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, 2015, pp. 39–48.
[30] N. An, Y. Xiao, J. Yuan, Y. Jiaoyun, G. Alterovitz, Extracting causal relations from the literature with word vector mapping, Comput. Biol. Med. 115 (2019) 103524, https://doi.org/10.1016/j.compbiomed.2019.103524.
[31] A. Akbik, J. Broß, Wanderlust: Extracting semantic relations from natural language text using dependency grammar patterns, in: www workshop, vol. 48, 2009.
[32] 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, in: Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2010, pp. 33–38.
[33] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, 2013, pp. 3111–3119.
[34] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv: 1810.04805 (2018).
[35] D.A. Lindberg, B.L. Humphreys, A.T. McCray, The unified medical language system, Methods Inform. Med. 32 (4) (1993) 281.
[36] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucl. Acids Res. 32 (suppl_1) (2004) D267–D270.
[37] S. Bird, Nltk: the natural language toolkit, in: Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, 2006, pp. 69–72.
[38] C.D. Manning, M. Surdeanu, J. Bauer, J.R. Finkel, S. Bethard, D. McClosky, The stanford corenlp natural language processing toolkit, in: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, 2014, pp. 55–60.
[39] R. Speer, J. Chin, C. Havasi, ConceptNet 5.5: An open multilingual graph of general knowledge, 2017, pp. 4444–4451. URL http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972.
[40] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, A. Joulin, Advances in pre-training distributed word representations, in: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), 2018.
[41] N. Reimers, BERT NLI Models, https://github.com/UKPLab/sentence-transformers/blob/master/docs/pretrained-models/nli-models.md [Online; accessed 20-April-2020].
[42] 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 (2019).
[43] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks (2019). arXiv:1908.10084.
[44] D. Chicco, G. Jurman, The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation, BMC Genom. 21 (1) (2020) 1–13.
[45] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, H. Pfister, Upset: Visualization of intersecting sets, IEEE Trans. Visual Comput. Graphics 20 (12) (2014) 1983–1992.
[46] M.-H. Hsu, M.-F. Tsai, H.-H. Chen, Query expansion with conceptnet and wordnet: An intrinsic comparison, in: Asia Information Retrieval Symposium, Springer, 2006, pp. 1–13.