MediCause: Causal Relation Modelling and Extraction from Medical Publications

Ioannis Reklos and Albert Meroño-Peñuela

King's College London, UK
{ioannis.reklos, albert.merono}@kcl.ac.uk

Abstract. Causal relations are one of the most important types of information that can be extracted from medical publications. Therefore, the automated extraction of such relations from medical text, and the development of ontologies for representing them, are active fields of research. Causal relation extraction is typically decomposed into causal sentence detection, entity recognition, and relation extraction. This study addresses the entity recognition sub-task, which remains largely unsolved since existing ontological models do not capture the various entities involved in causal relations, and datasets annotated with such entities are missing. Therefore, here we propose MediCause, an ontological model for entities involved in causal relations, and a novel dataset using it to annotate 1,202 causal sentences from existing datasets. We evaluate MediCause by training various BERT models that can recognize and label the entities in unseen texts, and we find that a BioBERT-large model fine-tuned with our dataset is the best model at this task (macro-averaged F1-score of 0.844). We also use MediCause to annotate entities in causal sentences from unseen, recent publications, and have experts evaluate them with encouraging results.

Keywords: Causal Relation Extraction · Ontology · Medicine · NLP

1 Introduction

One of the most important types of information that can be extracted from medical publications is causal relations, because they reveal the actions that need to be performed to produce a desirable outcome, such as curing a disease [15]. Once extracted, these relations can be used by automated systems and ontologies, such as [12,35,13,4], to produce e.g. causal knowledge graphs that structure and formalise valuable scientific knowledge and hypotheses for research [9,32]. However, in these approaches relations are typically manually annotated, and are represented either as text (limiting their processing) or as structured statements (limiting their complex, medical expressiveness). Moreover, in the medical domain these relations already exist in natural language, with more than one million papers added to the PubMed database each year [20]; this means, however, that manual annotation of causal relations is not feasible. Therefore, a viable solution is to automatically extract the relations and the roles of the entities involved from the texts.

Previous work addresses this task using rule-based and machine learning methods, with varying degrees of success [43]. More recently, state-of-the-art performance has been produced by fine-tuning pre-trained Bidirectional Encoder Representations from Transformers (BERT) [7] models to perform the relation extraction task. Although this approach has produced very good results [34,16], its application to the medical domain is very limited due to the lack of annotated causal sentences from medical publications for training [40]. Consequently, researchers have decomposed the causal relation extraction problem into three simpler tasks: (i) causal sentence detection; (ii) detection of entities involved in the causal relation; and (iii) causal relation extraction [19]. BERT models achieve high performance at causal sentence detection using small datasets of annotated causal sentences [19,44]. However, to the best of our knowledge no ontological model nor dataset exists with enough expressiveness and examples as to capture the entities involved in the causal relations, such as ”Increased BMI is associated with reduced life expectancy”, where the existing ontologies would either represent it as text or the triple (BMI, associated with, life expectancy), missing the ”increased” and ”reduced” which are vital to the relation. As a result, the application of state-of-the-art models like BERT to the task of recognizing the entities involved in the causal relations has been very limited.

In this paper we propose MediCause, an ontological model and a novel dataset for causal relation extraction. The ontological model addresses the limitations of the existing ontologies by capturing the causal relations as well as the roles of the entities involved in them; and the dataset contains 1202 causal sentences from the datasets of [19,44] with entities involved in the causal relations that we annotate with the ontological model. With these, we develop a pipeline of BERT models: one that detects sentences containing causal relations, and another that annotates the entities involved in the relations by assigning them role labels according to the MediCause ontology. Finally, we train various BERT models and compare their performance on the entity recognition task. We perform a quantitative and an expert-based qualitative evaluation for recent publications, achieving a macro-averaged F1-score of 84.4%, and the expert evaluators completely agreed with the information extracted by the model in more than 68% of the sentences annotated. Therefore, the research questions we address in this paper are: *How can an ontological model represent the various entities involved in causal relations? How effective are pre-trained scientific language models at the task of recognising the various entities involved in causal relations in medical texts?* The specific contributions of the paper are:

- MediCause, a novel ontological model for causal relations with causes, effects and cause and effect variables and specifiers, for the annotation of causal sentences; and a novel dataset with 1,202 annotated examples (Section 3)
- Experiments (Section 4) and results on using state-of-the-art language models, such as BioBERT, and standard baselines, such as SVM, at the task of learning how to extract causal sentence elements according to the roles defined by MediCause from data annotated with it

2 Related Work

Several approaches have been developed in Knowledge representation aiming to store scientific knowledge to facilitate information retrieval and hypothesis generation. Indeed, nanopublications have been developed to formally represent units of scientific information in the form of statements, which are represented as triples of the form `(subject,predicate,object)` [13]. Similarly, the Automated Discovery of Scientific Knowledge Ontology (DISK) provides a formal framework for representing scientific hypotheses [12]. An important limitation of these ontologies is that the causal relation (or hypothesis) is represented either as simple text [4,3] or as structured statements [12,13]. Both types of representations have drawbacks because textual statements limit the processing ability of machines as there is no information regarding the entities and their role in the relation [35]. At the same time, the structured representations lack the capacity to represent complex hypothesis statements because they consist of simple triples. This representation is far too simplistic to represent causal relations in the medical field. For example, the statement *"Increased BMI is associated with reduced life expectancy"* cannot be accurately represented by the triple `(BMI, associated with, life expectancy)`. An ontological model of causal relations that has the expressive ability to capture causal relations in medicine is the Simplified Biological Expression Language (SBEL) statement format: `(1st function, 1st entity, relation, 2nd function, 2nd entity)` [34]. The main limitation of this model is that due to being developed for use on BEL statements it cannot represent contextual information that specify the relationship, such as the age group or comorbidities of the patients involved in the studies. This is assumed to be provided by the BEL micropublications as experimental context. As such, a more expressive ontological model is required to fully capture these relations and structure them in a way that enables machines to efficiently reason with them. Finally, [35] represents research hypotheses in computable form and uses them to automatically formulate and test new hypotheses, but it is limited to processing Yeast metabolic pathways.

Another major issue with the above approaches is that the hypotheses must be manually annotated and input into the system. To tackle this problem, there is ongoing research aiming to develop automated systems that can extract hypotheses, as causal relations of the form (cause – causal connective – effect), from scientific text. Extracting causal relations from text can be broken down into three main sub tasks, namely causal sentence detection, recognition of entities involved in the relation, and relation extraction [40]. Recent research suggests that decomposing the causal relation extraction into these three sub-tasks can reduce the computational complexity of the process [19]. Given the complexity of detecting explicit and implicit causal relations, at the time of writing, most of the research has been attempting to solve the simpler task of extracting intra-sentential causal relations [45,19,23,39], which are causal relations contained in a single sentence, with fewer researchers attempting to develop systems that can extract inter-sentential relations [18,38], which are causal relations than extend to more than one sentence [43].The techniques used to solve these tasks can be divided into rule-based and machine learning methods.

Rule based relation extraction methods rely on a set of rules that need to be created and tuned by humans, which is a very time consuming process. Moreover, they may exhibit low performance due to ambiguity and usually only work on the specific domain they were created for, failing to generalize to other domains [23]. Indeed, [17] used a very small dataset consisting of 130 abstracts and achieved an F1 score of 68% on identifying causal relationships but only achieved an F1 score of 49.7% in identifying the cause and 51.2% in identifying the effect. Meanwhile, [2] focused on the very narrow area of Human Immunodeficiency Virus (HIV) drug resistance, and used a predefined list of mutations and drugs which means that even though their approach achieved an F1 score of 84.5% in relation extraction, the system would not be able to generalize outside the very narrow scope of the paper.

In contrast, machine learning methods, and deep learning in particular, automatically learn the features needed to perform the task. State of the art results in the causal sentence detection task have been obtained by fine tuning BERT models. BERT models have been used by [19] to classify sentences as causal, achieving an F1 score of 87% on the largest biomedical dataset tested in that study. In [44], the authors created a manually annotated dataset of causal sentences and compared the classification performance of Linear Support Vector Machine (SVM)s with Deep Learning approaches. The BioBERT model [21] performed best and achieved an F1 score of 88.1%. Deep learning approaches used to address the problem of causal relation extraction include Convolutional Neural Network (CNN) [23], Transformers [38], BERT [16,34] and Long Short-Term Memory (LSTM) [41]. The performance of CNN models of [23] varied depending on the dataset used for training and achieved a 91.82% F1-score on the International Workshop on Semantic Evaluation 2010 (SemEval-2010) task 8 dataset. The transformer model developed by [38] managed to achieve an F1 score of 66.2% in extracting both intra- and inter-sentential causal relations. Furthermore, [16] used BERT models and achieved an F1-score of 97.68% on the SemEval-2010 task 8 dataset. Finally, [34] used BioBERT models to extract the BEL statements in the BioCreative-V BEL Track corpus [10] and achieved state-of-the-art performance producing an F1-score of 54.8% in extracting the entire BEL statement.

The only dataset that is widely available and which is most often used as a benchmark for the task of causal relation extraction, is the SemEval-2010 task 8 dataset [14]. This dataset has significant drawbacks. Specifically, the cause-and-effect sentences contained therein are predominantly short and explicit [19], which is not representative of text which would be found in medical publications. Moreover, the sentences are not sourced exclusively from medical publications so the terminology and language used would not be the one needed for this study. Finally, the annotation of the relations is very poor, with only the cause and effect being annotated, which means that important information of the relation is discarded as irrelevant.

3 MediCause: Ontological Model and Dataset

3.1 Causal Sentence Detection

The dataset used in this study is a combination of the bio-causal dataset developed by [19] and the Detecting Causal Language in Science dataset (DCLS) developed by [44]. These datasets contain causal and non-causal sentences extracted from medical publications. The main difference between the two datasets is that the bio-causal dataset contains sentences that are annotated as causal (1) or non-causal (0), whereas the DCLS dataset further divides the sentences into "no relation" (0), "direct causal" (1), "conditional causal" (2) and "correlational" (3). To combine these two datasets, it was necessary to convert them into a common format. As such, given that in this work we are interested in extracting all kinds of causal relations between variables, all non-zero labels of sentences in the DCLS dataset were converted to 1 and then the two datasets were combined. A summary of the characteristics of the individual datasets and the combined dataset can be seen in Table 1.

Table 1. The characteristics of the datasets

Dataset	Causal Sentences	Non-Causal Sentences	Total
Bio-causal	1113	887	2000
DCLS	1705	1356	3061
Combined	2818	2243	5061

Pre-processing There are several pre-processing steps that are used in NLP to improve performance of the models. These steps include, but are not limited to, lowercasing, stemming and/or lemmatization, as well as removal of stop words. It has been established that BERT models that are trained on cased text (both uppercase and lowercase letters) perform better on Named Entity Recognition tasks, because the entities usually start with an uppercase letter [21]. Furthermore, in both [19,44] stemming and lemmatization was not performed as it may remove valuable information. Finally, it has been shown that removal of stop words does not benefit BERT models [29]. Based on the above, the only preprocessing performed on the text was to remove all characters that should not appear in biomedical texts, using regular expressions.

3.2 Entity Recognition

Having extracted the sentences that contain causal relations, the next step is to develop a model that can detect and label the entities involved in the relation. Through the background research, it became apparent that no complete ontological model for representing these relations exists for the medical domain and, correspondingly, no annotated dataset of such sentences exists [40]. Furthermore, given that the terminology and causal markers used in different disciplines shows limited overlap [27], using datasets from non-medical domains was not a viable alternative. As such, Medicause, a novel ontological model and the corresponding annotated dataset were developed for this study.

3.3 Hypothesis model

Due to the inherent representational limitations of the available ontologies, the first step prior to performing the annotation was to define a simple ontological model of causal relations.

Causal relations can be approached as relations among events and the corresponding variables/entities [26]. Furthermore, researchers aiming to separate causal from correlational relations refine the above approach by examining the relation in terms of independent and dependent variables [5]. Given that medical research aims to test whether some manipulation of one entity produces some effect on another, the latter approach was more appropriate. As such, based on [5,34] here we define a **causal relation to include a manipulation (cause), the independent variable (causal variable), the connective, the effect and the dependent variable (effect variable).**

One issue we encountered with the above definition is that in pure cause and effect relationships of the form "A causes B" there is no manipulation of the independent variable and no effect on the dependent one. To solve this problem, we extend the definition to include an implicit "Occurrence" in both the cause and the effect. In this way, when the relationship is of the form "Manipulation of A results in Effect in B" it becomes "Occurrence of Manipulation of A results in Occurrence of Effect in B" which does not alter the meaning of the relation. Additionally, when the relationship is of the form "A causes B" it becomes "Occurrence of A causes Occurrence of B" which solves the problem since "occurrence" is both the manipulation (cause) and the effect and does not alter the meaning of the relation. This extension in the definition of causal relation allows it to capture both causal and correlational relationships between entities.

Another important issue is that causal relations in medical texts may require information outside of the relation to capture their full meaning. For example, even in the simple sentence "Suicide is one of the leading causes of death among pre-adolescents and teens" contained in the SemEval-2010 dataset, suicide is annotated as the cause and death is annotated as the effect, totally ignoring the words "among pre-adolescents and teens" which specify the relation. As such, if one was to extract the entities annotated in the dataset, the resulting relation would be ((Suicide),(Death)) which is misleading. Therefore, our definition was further extended to include cause specifiers and effect specifiers, which are information that define and specify the cause and effect respectively. The final form of a causal relationship in MediCause is shown in Figure 1.

It should be noted that this is a formal definition of an explicit causal relation and that some of these terms in causal relationships in medical texts may be omitted due to them being implicit. e.g, see Figure 2.

(Occurrence), (Cause), (Causal Variable), (Causal Specifier)

↓

(Connective)

↓

(Occurrence), (Effect), (EffectVariable), (EffectSpecifier)

Fig. 1. The form of a causal relationship in MediCause.

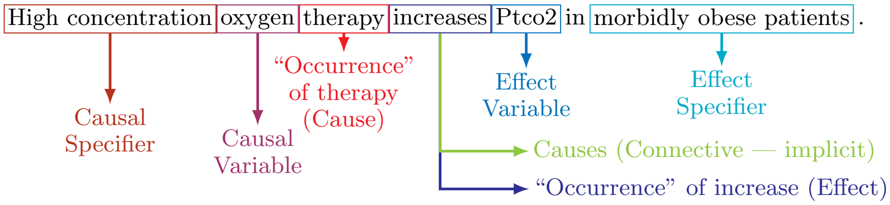


Fig. 2. Implicit connective (causes)

A comparison between the information identified by MediCause compared to the existing annotation of sentences of the SemEval-2010 dataset can be seen in Table 2.

Table 2. Difference in entities annotated by our model and the existing annotation of SemEval 2010. The labels shown are Causal Variable (VC), Connective (CON), Variable Effect (VE), Effect Specifier (ES), Effect (EF), Causal Specifier (CS), Variable Causal (VC) and Cause (C).

	the	chronic	inflammation	in	the	distal	part	of	the	stomach	caused	by	Helicobacter	pylori	inflection
SemEval			e1												e2
Extracted Information	((infection),(inflammation))														
Ours			EF			ES	ES			VE	CON		VC	CS	C
Extracted Information	((infection),(Helicobacter),(pylori),(caused),(inflammation),(stomach),(distal,part))														

In order to simplify the very complex task at hand, we assume that the cause, effect, and causal and effect variables are represented by one word, and that noun-phrases can be broken down into parts, as shown in Fig. 3.

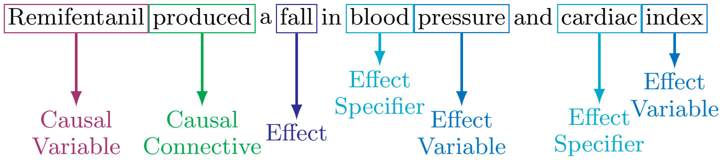


Fig. 3. Separation of the noun-phrase (blood pressure) into effect variable and specifier

Annotation. Training a machine learning model that can identify the entities involved in causal relations according to MediCause requires an annotated dataset of causal relations. Since such a dataset did not exist, one was manually annotated for the needs of this study. Specifically, the entities involved in the causal relations of 1202 causal sentences from the combined dataset of [19,44] were annotated according to the labeling rules mentioned above. The labels used for the annotation were developed using the Inside-Outside-Beginning (IOB) format [30], which is commonly used in token classification tasks, such as Named Entity Recognition (NER) [37].The labels used for the annotation can be seen in Table 3 and a sample annotation can be seen in Table 4.

Table 3. The labels used for the annotation of the dataset

Label	Entity
B-C	Cause
B-VC	Causal Variable
B-CS	Beginning Causal Specifier
I-CS	Inside Causal Specifier
B-CON	Beginning Connective
I-CON	Inside Connective
B-EF	Effect
B-VE	Effect Variable
B-ES	Beginning Effect Specifier
I-ES	Inside Effect Specifier
O	Outside

Table 4. Sample annotated sentence from the dataset

Endothelial	dysfunction	may	result	in	activation	of	platelets	and	coagulation	.
B-VC	B-C	O	B-CON	I-CON	B-EF	O	B-VE	O	B-VE	O

To create the MediCause Entity Recognition dataset, 1250 sentences from the combined datasets of [44] and [19] were randomly selected for annotation. To ensure that the annotation of the causal sentences would be correct, 50 randomly selected sentences from the MediCause dataset were annotated by both the first annotator and a second expert annotator who is a medical doctor and a researcher. The inter-annotator agreement was subsequently assessed by calculating Cohen's Kappa coefficient [6].The Cohen's Kappa value calculated between the two annotators in this study was 0.90, which is considered near perfect agreement [24]. After the annotation, most of the disagreements between annotators were resolved through discussion. A small number of disagreements between the two annotators could not be resolved, mainly due to the ambiguity of the text. The decision to not create stricter annotation rules that would eliminate the disagreements was made based on the work of [8], who state that developing very prescriptive annotation instructions to resolve all disagreements between annotators leads to the development of artificial datasets that do not represent the ambiguity present in natural language.

Given the high agreement between annotators and having resolved as many disagreements as possible, the remaining 1200 sentences were annotated by the first annotator. Of those sentences only 1152 were included in the MediCause dataset because the remaining 48 were misclassified as causal in the original datasets and did not contain a causal relation. The characteristics of the MediCause dataset, including the total count for each label, can be seen in Table 5.

Table 5. The characteristics of the MediCause Entity Recognition dataset

Label	**Count**
B-C	727
B-VC	1368
B-CS	1207
I-CS	849
B-CON	917
I-CON	940
B-EF	1184
B-VE	1684
B-ES	1914
I-ES	1570
O	15816
Total Tokens	28176
Total Sentences	1202

4 Evaluation

The MediCause dataset, models and code that we use in our evaluation are available online in GitHub and Zenodo.[1]

To evaluate our proposed MediCause ontological model and dataset, we follow a two-fold strategy with a quantitative and qualitative evaluation. For the quantitative evaluation, we assess the performance of transformer-based language models that we fine-tune with the annotated dataset at the tasks of causal sentence detection (which distinguishes sentences containing causal relations from those that do not) and entity recognition (which recognizes the various entities of our ontological model in causal sentences) with various metrics. For the qualitative evaluation, we ask medical experts about how much causal information the model captures, and how accurately, in unseen sentences that the model has previously annotated.

[1] https://github.com/IReklos/MediCause,
https://doi.org/10.5281/zenodo.6422025

4.1 Task: Causal Sentence Detection

To detect causal sentences, transfer learning was performed by using pre-trained BERT models and fine-tuning them on the combined dataset of [19,44] developed for this task. In addition to the original BERT models, we also fine-tuned SciBERT [1] and BioBERT [21] models which have been shown to produce state of the art performance on NLP tasks in scientific domains. After experimenting with different values for the epochs, learning rate and batch size based on the work of [7,36,31,19,44], the BERT models developed matched the performance reported by [19,44], with BioBERT(+PubMed + PMC) model achieving an F1-score of 0.881 when fine-tuned for 4 epochs with a learning rate of 2e-5 and a batch size of 32. The rest of the hyperparameters were kept the same as the pre-training of the models and we used Adam as the optimizer[2].

4.2 Task: Entity Recognition

To develop BERT multi-class classifiers that can recognize the entities involved in causal relations and compare their performance, we followed a similar process to the experimental design used for the causal sentence detection. For these experiments, we fine-tuned BERT models using a batch size of 32 and learning rates of 2e-5 and 3e-5. During initial testing, it became apparent that the number of training epochs needed for this task was larger and, based on initial results, 5,10 and 15 epochs were selected as the most suitable values. Moreover, we experimented with using a linear classification layer, as well as as a Multi-Layer Perceptron classifier consisting of a hidden layer of 500 non-linear units with ReLU activation function. The remaining training hyperparameters and the Adam optimizer parameters were kept the same. Given the large number of epochs required for fine-tuning, we hypothesized that since the dataset consists of the same sentences as the dataset used in the first task, the models that performed best on the first task would most likely perform best on the second task. Therefore, the models used in the Entity Recognition experiments were BioBERT(+PubMed), BioBERT(+PubMed + PMC), BioBERT-Large(+PubMed) and SciBERT with the custom vocabulary (scivocab). In addition to BERT, we also developed an SVM multi-class classifier using a one-versus-one approach [33], which serves as a baseline for the classification task. To train the SVM, the tokens were vectorized using FastText embeddings [25]. The best value of the hyperparameters of the SVM were determined by experimenting using both a linear and an RBF kernel, which is non-linear, and searching over different values of C using Grid Search.

As can be seen in Table 5, more than 56% of the tokens in the dataset are labeled as “O”, which makes it very unbalanced. This is a common issue in multi-class classification problems, with the distribution skew increasing with the number of classes [11]. To mitigate the resulting unbalance of classes as much

[2] Adam hyperparameter values: $\beta 1 = 0.9$, $\beta 2 = 0.999$, eps = 1e-6, weight decay of 0.01 and a linear warmup of 0.1.

as possible, class weights were calculated and used when calculating the error during model training to increase the weight of errors when classifying tokens labelled with the labels of interest and prevent the classifier from just predicting the majority class (“O”).

4.3 Metrics

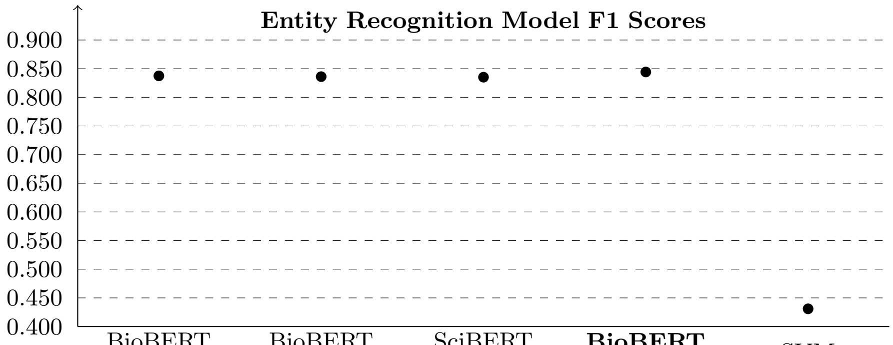
To evaluate the performance of the models we recorded their precision, recall and F1 score. These are the metrics proposed by [22] and used in all related work [19,44,34,23]. Given the relatively small size of the available datasets, the metrics were calculated using 5-Fold Cross-Validation which avoids splitting the dataset into training and test sets and thus produces a more accurate estimate of the performance of the models when trained on the entire dataset [42]. The evaluation of the performance of the models on the Entity Recognition task was done according the suggestions of [22]. The two methods that can be used to evaluate the performance of multi-class classifiers is micro- and macro-averaging of the scores for the individual labels.

Given that our dataset is very unbalanced and the performance of the models on the majority class (“O”) is the highest as shown in Table 6, micro-averaging would result in very high scores which would not accurately represent the performance of the models. As such, the macro-averaging method was used to produce the macro-averaged Precision, Recall and F1-score of the multi-class classifiers and the models were compared based on the macro-averaged F1-score.

Expert Evaluation To qualitatively evaluate the performance of the model and the expressive ability of the MediCause ontological model on unseen data, causal sentences from recent medical publications were extracted and annotated by our models. The resulting extracted relations were then graded by 5 medical doctors who evaluated whether all the information related to the causal relation was captured and whether the model identified the cause and effect correctly. The evaluation was done using a Likert scale ranging from ”Completely agree” to ”Completely disagree”. Moreover, in order to ensure that the evaluators could identify the cause and effect correctly, we added two of the manually annotated sentences as controls where we switched the cause and effect.

4.4 Results

In the Entity Recognition task, the best performance was produced by BioBERT-large which achieved a macro-averaged F1-score of 0.844 when trained for 15 epochs with a learning rate of 3e-5 and an MLP final classifier. All three of the remaining BERT models produced similar macro-averaged F1-scores ranging from 0.835 to 0.837 when trained for 15 epochs with a learning rate of 3e-5 and an MLP final classifier. Finally the SVM classifier achieved a macro-averaged F1-score of 0.431 when using an RBF kernel and $C = 5$. A comparison between the top F1-scores achieved by each model, as well as their macro-averaged precision and recall scores, can be viewed in Figure 4. Additionally, the Precision, Recall and F1-score produced by the best performing BioBERT-large model for each label are shown in Table 6.



	BioBERT (PD)	BioBERT (PD+PMC)	SciBERT (sci)	**BioBERT large**	SVM
Precision	0.813	0.814	0.813	0.826	0.395
Recall	0.866	0.862	0.860	0.864	0.521
F1 Score	0.837	0.836	0.835	**0.844**	0.431

Fig. 4. The highest macro-averaged F1-score produced by each model. The highest F1-score is shown in **bold**. The macro-averaged Precision, Recall and F1-score for each model are displayed as columns. For BioBERT models, PubMed is abbreviated to PD and for SciBERT models scivocab is abbreviated to sci.

Table 6. The average Precision, Recall and F1-score with standard deviation (±) over 5 folds produced by the best performing BioBERT-large model for each label.

Label	Precision	Recall	F1 Score
B-C	0.792 (±0.040)	0.846 (±0.024)	0.818 (±0.029)
B-CON	0.866 (±0.015)	0.924 (±0.017)	0.894 (±0.014)
B-CS	0.740 (±0.009)	0.817 (±0.012)	0.777 (±0.009)
B-EF	0.875 (±0.013)	0.895 (±0.017)	0.884 (±0.003)
B-ES	0.814 (±0.016)	0.866 (±0.017)	0.839 (±0.009)
B-VC	0.818 (±0.022)	0.836 (±0.029)	0.826 (±0.019)
B-VE	0.864 (±0.016)	0.902 (±0.013)	0.883 (±0.011)
I-CON	0.887 (±0.012)	0.954 (±0.013)	0.920 (±0.008)
I-CS	0.690 (±0.044)	0.741 (±0.043)	0.713 (±0.027)
I-ES	0.787 (±0.029)	0.813 (±0.049)	0.799 (±0.030)
O	0.955 (±0.006)	0.914 (±0.008)	0.934 (±0.003)

In the expert evaluation, in 68% of sentences the experts completely agreed that the model captured all of the information of the causal relation, Moreover, the cause and effect identified by the model was evaluated to be completely correct by the experts in 78% and 86% of the sentences respectively. It should be noted that there was only one ”Completely Disagree” response in identifying the effect and one in identifying the relation. The results of the expert evaluation are shown in Table 7.

Table 7. Results of the expert evaluation of the MediCause models.

	Completely Disagree	Somewhat Disagree	Not Agree or Disagree	Somewhat Agree	Completely Agree
Cause	1	0	0	10	39
Effect	0	0	1	6	43
Relation	1	0	1	14	34

4.5 Discussion

In the Entity Recognition task, the best performance was produced by the BioBERT-large multi-class classifier, which achieved a macro-averaged F1-score of 0.844. Furthermore, the SciBERT(scivocab) multi-class classifier produced a macro-averaged F1-score of 0.835 which was the lowest among the BERT models tested. The best results for all classifierrs were obtained for the labels"B-CON" and "I-CON" where BioBERT-large achieved F1-scores of 0.894 and 0.920 respectively. The very high scores achieved for those two labels can be attributed to the fact that each domain uses specific causal markers and one of the most important causal markers are the causatives (or causal verbs), with the number of different causatives used in each domain being fairly small [27]. Given that the "B-CON" and "I-CON" labels correspond to the causatives, identifying whether a word is part of a causative is a simple task, which explains the high F1-scores achieved by all models.

Interestingly, the baseline SVM multi-class classifier only managed to achieve a macro-averaged F1-score of 0.431 in the entity recognition task, even though its inputs were vectorized using FastText embeddings which are much more informative than the TF-IDF and Bag-of-Words vectors used in the Causal Sentence Detection task by [19,44]. The difference between the F1-score achieved by the SVM and the score achieved by BioBERT-large is 0.413, with BioBERT-large showing a performance improvement of 95.8% compared to the baseline. This difference not only highlights the complexity of the Entity Recognition task, which cannot be solved without providing information about the context of words, but also demonstrates the amount of information captured in the contextual embeddings produced by BERT. Furthermore, the increased complexity of this task allowed the BioBERT-large model to surpass the F1-score produced by all other models by 0.007 to 0.009. While these differences only represent an increase of 0.8 to 1% over the scores of the smaller BERT models, these results are in accordance with the differences in F1-score between BERT-base and BERT-large recorded by [7]. Even though the difference in the F1-scores produced by BioBERT-large and the smaller BERT models is small it was consistent across all hyperparameter combinations tested in this study, which suggests that it is the result of the improved embeddings produced by BioBERT-large compared to the other BERT models and cannot be attributed to the random initialization of the weights of the classification layer.

Finally, it is important to discuss the significance of the high F1-scores achieved by the BioBERT-large multi-class classifier in the Entity Recognition task in relation to the MediCause ontological model and dataset developed in this study. MediCause consists of two main parts, the development of an ontological model of the causal relations and the annotation of the entities in 1202 sentences. For the first part, the main considerations were to develop a definition which captures all the entities involved in causal relations, including relevant information that specify the relationship, while keeping the definition simple enough so that the underlying patterns can be learned by machine learning models trained on limited amounts of data. For the second part, the main consideration was to ensure that the annotation was as consistent as possible so that the machine learning models could extract the patterns without too much noise introduced by annotation errors. The performance of the BioBERT-large multi-class classifier developed using this dataset, suggests that the ontological model was sufficiently simple, allowing the BERT model to learn the underlying patterns and successfully classify unseen instances. Moreover, these scores suggest that the annotation of the dataset was consistent and did not introduce too much noise in the data which could have negatively impacted the performance of the models, especially given the small size of the dataset.

The expert evaluation of the performance of the model on unseen causal sentences shows that the more expressive MediCause ontology is able to capture all the information of the causal relations in medical texts. It is noteworthy that the experts agreement followed the F1-scores produced during model evaluation, with the experts completely agreeing with the effect in 86% of the sentences (0.88 F1-score) and the cause in 78% of the sentences (0.82 F1-score). The one sentence where one of the evaluators completely disagreed with the relation extracted by the model was "We found that among individuals without diabetes or other traditional ASCVD risk factors, there is an increased risk of incident CVD with increasing abnormal FPG levels." [46] where the model annotated "increasing abnormal FPG levels" as the cause and the evaluator argued that the cause was "individuals without diabetes". In that sentence the evaluator also did not agree or disagree with the effect annotated by the model (increased risk of incident CVD). The rest of the evaluators either "somewhat agreed" or "completely agreed" with the annotations made by the model on this sentence. In other sentences where the evaluators "somewhat agreed" or "did not agree or disagree" with the model, the model correctly annotated the cause and effect but missed some of the contextual information, as is the case with the phrase "CMR-FT derived GLS is a powerful independent predictor of MACE in patients with HCM" [28] where the model missed the word "derived" and just annotated "GLS" as the causal variable and "CMR-FT" as a causal specifier.

5 Conclusion

The purpose of this study was to address the unsolved problem of recognizing the entities involved in causal relations in medical publications. According to recent research, the proposed solution to this problem is to first detect causal sentences and then recognize of the entities involved in the causal relations of those sentences. For the detection of causal sentences, the state of the art results of [19,44] were replicated and the produced BERT model is able to detect causal sentences with very high accuracy. For the entity recognition task, we first developed the MediCause ontological model of causal relations in medical text, which addresses the limitations of existing ontological models, and used it to manually annotate the entities of 1202 causal sentences. This dataset was then used to fine-tune BioBERT and SciBERT models that can recognize the entities involved in causal relationships and label them, achieving very high performance on the task, as can be seen by both the metrics and the expert evaluation. Answering the research questions, compared to previous ontological models, MediCause effectively captures the entities involved in causal relations and represents the relations in a fine-grained way which can enhance the ability of machines to parse it and extract new information and research hypotheses. Moreover, the models developed using the novel dataset developed in this study are very effective in recognising the entities involved in causal relations and can be used to extract and annotate causal sentences from medical publications or annotate the entities of causal relations stored in textual format in existing ontologies.

We plan on extending our work in various ways. First, since one of our main limitations is the relatively small size of the MediCause dataset, we plan on extending the annotated sentences to improve performance, especially at recognizing Causal Specifiers and Effect Specifiers. Second, we will improve both the architecture of the final classifier for the BERT models –by using a CNN to extend token embeddings with their weighted neighborhoods– and the hyperparameter space search –by increasing number of epochs and learning rate. Third, we will continue with the natural step of extracting the causal relations between multiple entities beyond pairs in different relations, effectively building a knowledge graph of interconnected causal relations. Such a knowledge graph could open various possibilities for inference in scientific settings, by using it for e.g. generating graph embeddings (for causal link discovery and completion), reasoning (for consistency checking and derived causes and effects), hypothesis generation techniques, etc. Finally, we plan on experimenting with the methods we use in this paper for causal relation extraction in other areas of science beyond medicine, in order to investigate their generality.

References

1. Beltagy, I., Lo, K., Cohan, A.: SciBERT: A pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3615–3620. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1371, `https://aclanthology.org/D19-1371`
2. Bui, Q.C., Nualláin, B.Ó., Boucher, C.A., Sloot, P.M.: Extracting causal relations on HIV drug resistance from literature. BMC Bioinformatics **11**(1), 1–11 (Feb 2010). https://doi.org/10.1186/1471-2105-11-101

3. Ciccarese, P., Wu, E., Wong, G., Ocana, M., Kinoshita, J., Ruttenberg, A., Clark, T.: The swan biomedical discourse ontology. Journal of biomedical informatics **41**(5), 739–751 (2008)
4. Clark, T., Ciccarese, P.N., Goble, C.A.: Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications. Journal of biomedical semantics **5**(1), 1–33 (2014)
5. Cofield, S.S., Corona, R.V., Allison, D.B.: Use of causal language in observational studies of obesity and nutrition. Obesity Facts **3**(6), 353–356 (2010). https://doi.org/10.1159/000322940
6. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**(1), 37–46 (Apr 1960). https://doi.org/10.1177/001316446002000104
7. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT 2019. pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2018)
8. Dumitrache, A., Aroyo, L., Welty, C.: Crowdsourcing ground truth for medical relation extraction. ACM Trans. Interact. Intell. Syst. **8**(2) (Jul 2018). https://doi.org/10.1145/3152889
9. Färber, M.: The microsoft academic knowledge graph: a linked data source with 8 billion triples of scholarly data. In: International Semantic Web Conference. pp. 113–129. Springer (2019)
10. Fluck, J., Madan, S., Ansari, S., Kodamullil, A.T., Karki, R., Rastegar-Mojarad, M., Catlett, N.L., Hayes, W., Szostak, J., Hoeng, J., Peitsch, M.: Training and evaluation corpora for the extraction of causal relationships encoded in biological expression language (BEL). Database **2016** (2016). https://doi.org/10.1093/database/baw113
11. Forman, G.: An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research **3**, 1289–1305 (Mar 2003)
12. Garijo, D., Gil, Y., Ratnakar, V.: The DISK hypothesis ontology: Capturing hypothesis evolution for automated discovery. In: K-CAP Workshops. pp. 40–46. Austin, TX, USA (Dec 2017)
13. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Information Services & Use **30**(1–2), 51–56 (Sep 2010). https://doi.org/10.3233/ISU-2010-0613
14. Hendrickx, I., Kim, S.N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., Szpakowicz, S.: SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In: Proceedings of the 5th International Workshop on Semantic Evaluation. pp. 33–38. Association for Computational Linguistics, Uppsala, Sweden (Jul 2010), `https://aclanthology.org/S10-1006`
15. Karhausen, L.R.: Causation: the elusive grail of epidemiology. Medicine, Health Care and Philosophy **3**(1), 59–67 (2000). https://doi.org/10.1023/a:1009970730507
16. Khetan, V., Ramnani, R.R., Anand, M., Sengupta, S., Fano, A.E.: Causal-BERT : Language models for causality detection between events expressed in text. CoRR **abs/2012.05453** (2020), `https://arxiv.org/abs/2012.05453`
17. Khoo, C.S.G., Chan, S., Niu, Y.: Extracting causal knowledge from a medical database using graphical patterns. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics. pp. 336–343. Association for Computational Linguistics, Hong Kong (Oct 2000). https://doi.org/10.3115/1075218.1075261, `https://aclanthology.org/P00-1043`

18. Kruengkrai, C., Torisawa, K., Hashimoto, C., Kloetzer, J., Oh, J.H., Tanaka, M.: Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31. PKP Publishing Services Network (Feb 2017), `https://ojs.aaai.org/index.php/AAAI/article/view/11005`
19. Kyriakakis, M., Androutsopoulos, I., i Ametllé, J.G., Saudabayev, A.: Transfer learning for causal sentence detection. In: Proceedings of the BioNLP 2019 workshop. pp. 292–297. Association for Computational Linguistics, Florence, Italy (Aug 2019)
20. Landhuis, E.: Scientific literature: Information overload. Nature **535**(7612), 457–458 (Jul 2016). https://doi.org/10.1038/nj7612-457a
21. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (09 2019). https://doi.org/10.1093/bioinformatics/btz682
22. Li, J., Sun, A., Han, J., Li, C.: A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering pp. 1–1 (2020). https://doi.org/10.1109/TKDE.2020.2981314
23. Li, P., Mao, K.: Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. Expert Systems with Applications **115**, 512–523 (Jan 2019). https://doi.org/10.1016/j.eswa.2018.08.009
24. McHugh, M.L.: Interrater reliability: the kappa statistic. Biochemia medica **22**(3), 276–282 (2012)
25. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A.: Advances in pre-training distributed word representations. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018), `https://aclanthology.org/L18-1008`
26. Mirza, P., Tonelli, S.: CATENA: CAusal and TEmporal relation extraction from NAtural language texts. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 64–75. The COLING 2016 Organizing Committee, Osaka, Japan (Dec 2016), `https://aclanthology.org/C16-1007`
27. Mulkar-Mehta, R., Gordon, A.S., Hobbs, J.R., Hovy, E.: Causal markers across domains and genres of discourse. In: Proceedings of the Sixth International Conference on Knowledge Capture. p. 183–184. K-CAP '11, Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/1999676.1999716
28. Negri, F., Muser, D., Driussi, M., Sanna, G.D., Masè, M., Cittar, M., Poli, S., De Bellis, A., Fabris, E., Puppato, M., et al.: Prognostic role of global longitudinal strain by feature tracking in patients with hypertrophic cardiomyopathy: The strain-hcm study. International Journal of Cardiology (2021)
29. Qiao, Y., Xiong, C., Liu, Z., Liu, Z.: Understanding the behaviors of BERT in ranking. CoRR **abs/1904.07531** (2019), `http://arxiv.org/abs/1904.07531`
30. Ramshaw, L.A., Marcus, M.P.: Text chunking using transformation-based learning. In: Text, Speech and Language Technology, pp. 157–176. Springer Netherlands (1999). https://doi.org/10.1007/978-94-017-2390-9_10
31. Rosvall, E.: Comparison of sequence classification techniques with BERT for named entity recognition. Ph.D. thesis, KTH Royal Institute Of Technology School of Electrical Engineering and Computer Science, Stockholm, Sweden 2019 (2019)

32. Salatino, A.A., Thanapalasingam, T., Mannocci, A., Osborne, F., Motta, E.: The computer science ontology: a large-scale taxonomy of research areas. In: International Semantic Web Conference. pp. 187–205. Springer (2018)
33. Sánchez-Marono, N., Alonso-Betanzos, A., García-González, P., Bolón-Canedo, V.: Multiclass classifiers vs multiple binary classifiers using filters for feature selection. In: The 2010 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (Jul 2010). https://doi.org/10.1109/ijcnn.2010.5596567
34. Shao, Y., Li, H., Gu, J., Qian, L., Zhou, G.: Extraction of causal relations based on SBEL and BERT model. Database **2021** (Jan 2021). https://doi.org/10.1093/database/baab005
35. Soldatova, L.N., Rzhetsky, A.: Representation of research hypotheses. In: Proceedings of the Bio-Ontologies Special Interest Group Meeting 2010, vol. 2, pp. 1–15. Springer Science and Business Media LLC (2011). https://doi.org/10.1186/2041-1480-2-s2-s9
36. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune BERT for text classification? In: Sun, M., Huang, X., Ji, H., Liu, Z., Liu, Y. (eds.) China National Conference on Chinese Computational Linguistics, vol. 11856, chap. Lecture Notes in Computer Science. Springer, Cham, Switzerland (Oct 2019). https://doi.org/10.1007/978-3-030-32381-3_16
37. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the conll-2003 shared task: Language-independent named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4. pp. 142—-147. CONLL '03, Association for Computational Linguistics, USA (2003). https://doi.org/10.3115/1119176.1119195
38. Verga, P., Strubell, E., McCallum, A.: Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 872–884. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-1080, `https://aclanthology.org/N18-1080`
39. Wang, L., Cao, Z., de Melo, G., Liu, Z.: Relation classification via multi-level attention CNNs. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1298–1307. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/P16-1123, `https://aclanthology.org/P16-1123`
40. Xu, J., Zuo, W., Liang, S., Zuo, X.: A review of dataset and labeling methods for causality extraction. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 1519–1531. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). https://doi.org/10.18653/v1/2020.coling-main.133, `https://aclanthology.org/2020.coling-main.133`
41. Xu, Y., Mou, L., Li, G., Chen, Y., Peng, H., Jin, Z.: 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. pp. 1785–1794. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015). https://doi.org/10.18653/v1/D15-1206, `https://aclanthology.org/D15-1206`
42. Yadav, S., Shukla, S.: Analysis of k-Fold Cross-Validation over Hold-Out Validation on colossal datasets for quality classification. In: 2016 IEEE 6th Interna-

tional Conference on Advanced Computing (IACC). pp. 78–83. IEEE (Feb 2016). https://doi.org/10.1109/iacc.2016.25

43. Yang, J., Han, S.C., Poon, J.: A survey on extraction of causal relations from natural language text. CoRR **abs/2101.06426** (2021), https://arxiv.org/abs/2101.06426
44. Yu, B., Li, Y., Wang, J.: Detecting causal language use in science findings. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 4656–4666 (2019), https://www.aclweb.org/anthology/D19-1473.pdf
45. Zhang, Y., Qi, P., Manning, C.D.: Graph convolution over pruned dependency trees improves relation extraction. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2205–2215. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1244, https://aclanthology.org/D18-1244
46. Zuo, Y., Han, X., Tian, X., Chen, S., Wu, S., Wang, A.: Association of impaired fasting glucose with cardiovascular disease in the absence of risk factor. The Journal of Clinical Endocrinology & Metabolism (2021)