Institut für Telematik
Pervasive Computing Systems / TECO
Leiter: Prof. Dr.-Ing. Michael Beigl

Document-Level Causality Extraction for Impact Evaluation

Bachelor´s Thesis

by

Ege Uzhan

Department of Informatics

Responsible Supervisor:	Prof. Dr. Michael Beigl
Supervising Staff:	Ployplearn Ravivanpong
Project Period:	30.11.2022 - 30.03.2023

Kurzfassung

Mit den Entwicklungen in der Technologie und vornehmlich im Bereich der Verarbeitung der natürlichen Sprache (NLP), Herausziehen der Informationen aus den enormen Daten der Internet ist durch die Zeit einfacher und kompetenter geworden. Ein spezieller Bereich der NLP, der noch eine schwere Aufgabe verursacht, ist die Kausalitätsextraktion, welche sich um die wichtige Beziehung von Kausalität kümmert, wobei Kausalität eine große Rolle in unserem Leben spielt. Aber trotz der sich vergrößenden Anzahl der Methoden zur Kausalitätsextraktion, die meisten Methoden dafür eignen sich auf die Relationen in der Satzebene. Die Kausalitätsextraktion in der Dokumentenebene bleibt noch als eine relativ ungefasste Aufgabe. Außerdem ist die Kausalitätsbeziehung aufgrund ihrer Natur stark domänenabhängig, die Mangel der annotierten Daten in spezifischen Domänen ist immer noch ein Problem. Als eine wichtige Domäne, es existieren keine Anwendungsfälle der Kausalitätsextraktion zur Analyse der Wirkungen von Entwicklungsprojekten und Politiken, diese werden auch Interventionen genannt. Unter Berücksichtigung von der höhen Anzahl der Informationen in Dokumentenforme und die Verfügbarkeit der Wirkungsevaluierungen von Interventionen als Dokumenten, wir schlagen ein System zur Extraktion der kausalen Beziehungen von Dokumenten durch Kombinieren von anderer Werke vor, wobei wir die Wirkungsevalierungen zur Interventionen als die spezifische und relativ unberührte Domäne zum Überprüfen des Systems nutzen. Wir untersuchen, wie die kausale Inferenz genutzt werden kann, um die Interventionen und ihre Wirkungen zu extrahieren, die von diese Wirkungsevalierungen überprüft wurden. Unser System hat ein Ansatz, der von den verfügbaren Modellen zur kausalen Inferenz unterschiedlich ist: wir extrahieren zuerst die möglichen Kandidaten der Ursache und Wirkung von einer möglichen kausalen Beziehung, und dann suchen wir für existierende Beziehungen unter der gefundenen Kandidaten. Wir untersuchen auch die Domänen der Datensätze, die wir zur Training nutzen, und fragen dann, ob die Datensätze über generelle Domänen der Kausalität genügend für das Extrahieren von der Kausalität in domänen-spezifischen Dokumenten von Wirkungsevaluierungen. Unsere Ergebnisse zeigen, dass das Erkennen von der Ursachen- und Wirkungskandidaten zur Entities nicht genügend in beiden Arten von Datensätzen ist, während die domänen-spezifische Datensatz relativ bessere Ergebnisse liefert. Indessen zeigt das Beziehungsextraktionsteil von unserem System versprechende Ergebnisse, falls der domänen-spezifische Datensatz von Wirkungsevaluierungen genutzt wird. Neue annotierten Daten für die kausale Extraktion mit einem einstimmig beschlossenen Kennzeichnungssystem, mehr annotierte Daten für unsere Domäne und eine Verbesserung vom Eigennamenerkennungsanteil des Systems wären unter möglichen zukünftigen Werke.

Abstract

With the developments in technology and the Natural Language Processing field especially, extracting information from the huge data available on the internet has become easier and more competent. One specific field that remains a challenge is causality extraction, which deals with the special relation of causality that actually plays a big role in life in general. Even with the increasing amount of methods aiming to extract causality, this has been mostly limited to the form of sentence-based extraction. The extraction of causality at the document level remains a relatively untouched task. Furthermore, by its nature, the causal relation is very domain-dependent, and the lack of labeled data for specific domains remains a challenge. As one significant domain, there is yet a use case for using causality extraction to analyze the cause and effect in development projects and policies, also interventions. Considering the high amount of information in document forms and the availability of impact evaluations of interventions as documents, we propose a framework for extracting causal relations from documents by combining several existing works, whereby we use impact evaluations as a specific and relatively untouched domain to inspect our framework. We study how causality extraction can be used to obtain the interventions and their effects studied by impact evaluations. Our framework takes a different approach from the available causality extraction models, where we first extract possible cause and effect candidates, and then search for possible relations between the found candidates. We also inspect the domains of datasets used to train the models we are using and ask, whether general domain datasets for causality are sufficient to extract causality from domain specific documents of impact evaluations. Our results show that the recognition of the cause and effect candidates is not sufficient with either of the datasets, while the results of the domain-specific dataset also delivers relatively better but still poor results. The results imply that relation extraction models are also not applicable on general data for the intervention and effect extraction, while they show promising results on the impact evaluation dataset. New labeled data for causal extraction with a labeling scheme that would be unanimously agreed upon, more labeled data for the field and an improvement for the Named Entity Extraction section are among possible future works.

Contents

List of Figures

List of Tables

Abbreviations

NLP	Natural Language Processing
NER	Named Entity Recognition
RE	Relation Extraction
Bi-LSTM	Bidirectional Long Short Term Memory
CRF	Conditional Random Field
IE	Impact Evaluation
3ie	International Initiative for Impact Evaluation[24]

1. Introduction

With the developments in technology over the last few decades, the amount of information available has increased immensely. The vast resources of data available creates a requirement of techniques for reaching this information in a shorter amount of time with less effort. As a task of Natural Language Processing (NLP), Information Extraction has a major role in the future of NLP regarding the vast increase of data [56]. Many different subcategories of information extraction have emerged in time, such as event extraction, relation extraction, and named entity recognition[56].

One important special field in NLP is the extraction of a special type of relationship that exists in every aspect of life: Causality. Causality, as a relation, acts as an unwritten law of our universe. Even at times that we are not aware of it, we are acting as a result of an event that has occurred before. Causality lays a foundation of the basics of reasoning for us [55]. We, as humans, notice this rule of nature as we grow up.

In the field of NLP, there are many domains which causal extraction has been applied on: It has been used in predicting possible events through news events [46]. In medicine, it was used to infer HIV drug resistance from medical literature [4] and to infer possible cures for specific diseases [28].

As seen from the works above, causality extraction has been receiving a fair amount of recognition in several domains in recent years. However, as far as we are aware, most works focus on the task of causality extraction in a sentence-level approach: The works aim to infer the causal relations inside a sentence, where in most cases every sentence contains an explicitly stated causal relation [56]. On the other hand, an enormous amount of data on the internet is available as documents. Scientific papers, articles, books, and even summaries are available in a document format. These huge chunks of text can contain unrelated information, or information that can be mostly disregarded for the needed part of information for the reader. Using information extraction on a document-level scale can reduce the time required to gather the necessary information. While the relation extraction task on documents has seen several touches [60], [62], [61], [54] in recent years, a document-level extraction specifically on causality has almost never been touched as a topic, as we are aware.

It should be noted that causality, and especially the causal markers in the causal relationship, are fairly dependent on specific domains [38]. There are several example datasets prepared for causal relations [14], [19], [9], and while some of these datasets are prepared with general examples of causality, several of them are focusing on a specific domain, like Adverse Drug Effects(ADE)[16] corpus does for medicine. There are also many domains lacking an exclusive dataset. A concrete domain that requires attention are the impact evaluations. An impact evaluation is a special evaluation that tries to infer the effect of an intervention on a resulting outcome[23]. As far as we are aware, there is only one work creating an intervention corpus from humanitarian assistance programms and proposing an approach for their extraction from texts [33], however, it is not publicly available.

Contribution: Considering the availability of impact evaluation data in forms of documents and how the domain of impact evaluation lacks labeled data, we want to apply the concept of document-level causal inference, while focusing on the field of impact evaluations and creating a dataset specific for the field. To evaluate the situation in the present works, we propose a framework for extracting causal relations from documents by combining several existing works, whereby we use impact evaluations to inspect our framework. By creating the framework, we will study if the given parts are able to extract the required information. Our work differs from the other work in its contribution and its methodology. The main contribution of our work is going to be the annotated dataset of impact evaluation summaries from the International Initiative for Impact Evaluation (3ie) [11], and our methodology aims to first extract cause-effect entity candidates, then it tries to infer causality between the extracted entities. A use of a similar approach in other works is unbeknownst to us. Our main research questions are as follows:

1. Is causal NLP a reliable method for extracting economic interventions from literature?
2. Can we find out which intervention is successful and which is not by using causality extraction?
3. Can the models which are trained on general data be applied to more specific and relatively untouched fields like economics and politics?

Structure of the Thesis: This bachelor thesis will be focusing on document-level causality extraction from impact evaluation reports. In the next chapter, main concepts required to dive into the topic will be presented, such as causal relations, named entity recognition, relation extraction and also details of the methods used in our framework. The third chapter explains our methodology, where we present our labeling scheme, the dataset created for the task and the pipeline for causality extraction. Later on in the fourth chapter, we present the experimental setup for our study, then show and discuss our results. The fifth and final chapter inspects our conclusions and discusses possiblities in future works.

2. Background & Related Work

In this chapter, necessary background information regarding causality, main tasks of our framework and the models used for our framework will be presented. Additionally, general information about impact evaluations and several related works similar to our task will be explained.

2.1 Causality

Causality (or a causal relation) can be defined as a "relationship between something that happens and the reason for it happening" [44]. Another description of causality would be a relation occurring between two events e_1 and e_2, where the event e_2 is the outcome of the event e_1 [7]. A common way to categorize causality is according to their syntactic existence: explicit causality and implicit causality [56]:

Explicit causality: This type of causality is given when the causal relation is accompanied by an explicit identifier in the text [56]. Khoo et al.[27] enumerate the following language constructs to express explicit causality[2]: *Causal links* like because, so, since[1]; *Causative verbs* like break and kill[51]; *resultative constructions*, *if-then conditionals* and *causation adverbs.* It should be noted that several explicit constructs can be ambiguous in expressing causality, such as the causal link *from* [7], on some occasions a connector known to indicate causality for one context might not be expressing a causal relation in another context.

Implicit Causality: In case of an implicit causality the theme of the text and the semantics have a huge role for inferring the causal relation. Hereby, the causal meaning from the explicit connectives are either expressed through ambiguous connectives, or they are not expressed by any explicit structure at all [56]. Khoo et al. emphasize that there are many cases of causal expressions through use of implicit causality and this must be inferred by the reader without explicit clues [27]. The following is an example of an implicit causal relation: "He left his bike outside without locking it properly. ... He could not find it next morning, it was stolen." Hereby, the theft of the bike is the result of the actor not locking his bike, but there are no explicit clues or phrases expressing the causal relation.

We can also classify causality regarding the locations of the cause and effect entities in the text: Intra-sentential causality is the case where the cause and effect entities for one causal relation are found within a sentence [56]. On the other hand, inter-sentential causality is the term for the occasion when cause and effect appear in different sentences [56]. In explicit cases, several spans like ”As a result” and ”This causes” can indicate an inter-sentential causality [3]. Another type worth mentioning is the chained causality [3], or a transitive causal relation. In this case the result of an action is also a cause of one other. This complex causality form also presents a challenge.

2.2 Causality Extraction

In the field of NLP; Causality Extraction, Causal Relation Extraction or Causal Inference can be described as the task of figuring out the existence of a causal relationship considering an effect's occurrence circumstances [58]. Task of causality extraction aims to extract causal relations between annotated entities [56], where the direction of the relation also plays a role in extraction. There are several approaches of doing causality extraction, the main three categories of approaches are knowledge-based(rule or pattern based) extraction, statistical machine learning based extraction and deep learning based approaches using neural networks [56].

We can describe the causality extraction task as a combination of two NLP tasks: Named Entity Recognition (NER) and Relation Extraction (RE).

2.2.1 Named Entity Recognition

According to Nadeau et al.[39], Named Entity Recognition (NER) is the task of recognizing mentions of entities which can be regarded as information units [39]. The original task is restricted into mentions of the *rigid designators*: proper nouns in most approaches [39]. Some example entity classes would be ”Person”, ”Organization”, ”Date” and ”Location” [39]. There are several approaches for the NER task: Rule-based approaches, unsupervised learning approaches, feature-based supervised learning approaches and deep-learning based approaches [30].

State-of-the-art approaches
This section explains one state-of-the-art method that we have considered for our experiment.

1. **Bi-LSTM-CRF[26][22][45]**:
   The Bi-LSTM-CRF deals with the task of NER with a different approach: *Sequence Labeling(Tagging).* Sequence Labeling is a special method in Natural Language Processing that aims to predict the category of labels for every morpheme in a text, where the labels imply that the morphemes have a similar role syntactically [17]. The Sequence Labeling method can be used for several

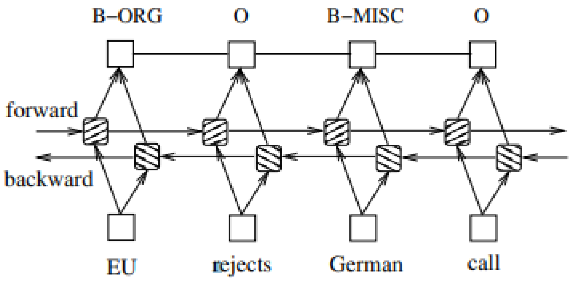


Figure 2.1: Bi-LSTM-CRF representation[22]

tasks of the Natural Language processing, such as Part-of-Speech Tagging, Named Entity Recognition and Text Chunking [17].

The Bi-LSTM-CRF model is a combination of two separate modules: the bidirectional LSTM [15] and CRF(Conditional Random Field) [29]. The bi-LSTM allows the past and future features to be extracted in two directions for each token, which are then used to create a global representation of the whole sequence [17]. The CRF added on top of the LSTMs allows the prediction of the tag information [22]. Bi-LSTM-CRF has been used in several works for causality extraction alone before [36], [7], but not as a part of a pipeline like our approach.

2.2.2 Relation Extraction

Relation Extraction (RE) is a task of extraction and classification of semantic relationships that were found from texts [49]. Like causality, there are many types of relations that can be extracted from text. Some example relations would be ”part_of”, ”country”, ”spouse”, ”educated_at” and ”publication_date” [59].

Our work will focus on one of the subtasks of the RE, namely the document-level relation extraction. Document-level extraction implies that the relations found in the text are not only limited to being inside a sentence, but any part of the relation can be found anywhere in a given document. As multiple sentences must be considered for a relations existence, sentence-level relation extraction is not sufficient and document-level relation extraction methods must be applied [59]. Entity coreferences and mentions of the same entity are also being considered [59].

Models

This subsection explains the state-of-the-art methods that we have considered for our experiment. The models we consider deal with Relation Extraction, as we could not find any relation extraction method that solely focuses on causal inference. The following model descriptions are based on their corresponding works [62], [61], [13], [54]. For the works, an entity pair should be considered as the pair (e_s, e_o), where e_s signifies the subject entity, and e_o stands for the object entity. For the following models, the following parameters are signifying the following: d: document length, e_i: entity $i \in \{s, o\}$ for the entity pair, $W_r \in \mathbb{R}^{d \times d}$: weight parameters for the relation, $W_i \in \mathbb{R}^{d \times d}$: weight parameters for the entity $i \in \{s, o\}$ [62], [54], [61], [13].

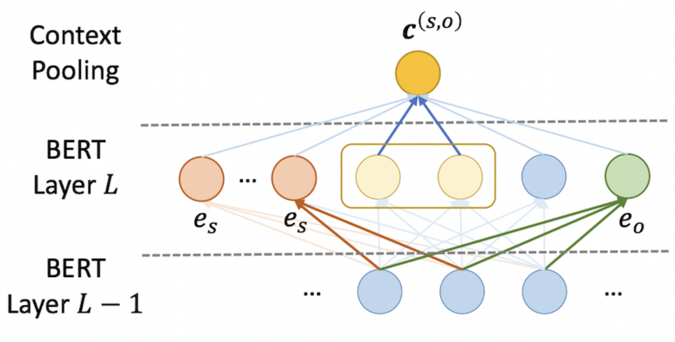


Figure 2.2: Localized Context Pooling[62]

1. **ATLOP**[62]:
   Proposed by Zhou et al.[62], ATLOP stands for "Adaptive Thresholding and Localized Context Pooling". The model introduces two new techniques of adaptive thresholding and localized context pooling for predicting the relations [62].

   ATLOP is introduced with two main steps: a modified BERT[8] Baseline with an Encoder and a Binary Classifier (incorporating the localized context pooling), and adaptive thresholding for calculating the loss [62].

   The process of prediction for ATLOP can be summarized as following: First, after the contextual embeddings H are extracted through a pretrained BERT module and pooling is performed to extract the entity embeddings h_{e_i} for all mentions of an entity, the model calculates the probability $P(r|e_s, e_o)$ of a relation r existing between the entities e_s and e_o through the following sigmoid activation [62]:

$$P(r|e_s, e_o) = \sigma(\sum_{j=1}^{k} z_s^{j\intercal} W_r^j z_o^j + b_r) \tag{2.1}$$

$$z_i = [z_i^1; ...; z_i^k] \tag{2.2}$$

   where b_r is the bias, and z_o^j and z_s^j represent the hidden states z_i split into k groups and W_i^j representing the respective parameters of the split. Hidden states are calculated with [62]:

$$z_i = \tanh(W_i h_{e_i}) \tag{2.3}$$

   The authors introduce the concept of localized context pooling, which is used to incorporate a local context embedding for each entity pair. Using the multi-head attention matrix of pre-trained models for each token, the localized context embeddings are calculated through the following [62]:

$$A^{(s,o)} = A_s^E . A_o^E \tag{2.4}$$

$$q^{(s,o)} = \sum_{i=1}^{H} A_i^{(s,o)} \tag{2.5}$$

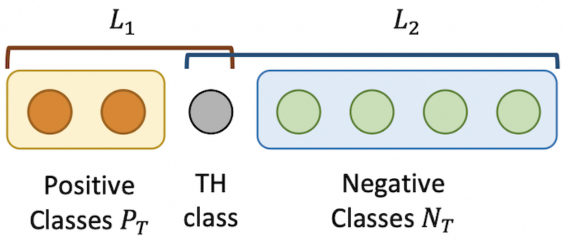


Figure 2.3: Adaptive Thresholding[62]

$$a^{(s,o)} = \frac{q^{(s,o)}}{1^{\intercal} q^{(s,o)}} \tag{2.6}$$

$$c^{(s,o)} = H^{\intercal} a^{(s,o)} \tag{2.7}$$

where A_i^E represents the entity-level attention matrix for an entity i in a pair, and $c^{(s,o)}$ represents the localized context embedding [62].

The localized context embeddings are then applied on the embeddings calculated in Equation (2.3) to get the final entity representations for each $i \in \{s, o\}$ [62]:

$$z_i^{(s,o)} = \tanh\left(W_i h_{e_i} + W_{c_1} c^{(s,o)}\right) \tag{2.8}$$

Then, the resulting probabilities are given to an adaptive thresholding module. The adaptive thresholding technique proposed by the authors mainly acts as a learnable threshold for the prediction probability $P(r|e_s, e_o)$. Adaptive thresholding (Figure 2.3) uses a class that acts as a thresholder between the sets of positive and negatively labeled classes for one entity pair. One class in the given figure represents a relation, if the relation exists between the given pair, it resides in P_T, else in N_T [62].

$$\mathcal{L}_1 = -\sum_{r \in P_T} \log\left(\frac{\exp(logit_r)}{\sum_{r' \in P_T \cup \{TH\}} exp(logit_{r'})}\right) \tag{2.9}$$

$$\mathcal{L}_2 = -\log\left(\frac{\exp(logit_r)}{\sum_{r' \in N_T \cup \{TH\}} exp(logit_{r'})}\right) \tag{2.10}$$

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 \tag{2.11}$$

A new loss for the adaptive thresholding is introduced ($\mathcal{L}$), where the $\mathcal{L}_1$ is calculated for the positive classes and $\mathcal{L}_2$ is calculated for the negative classes. $logit_r$ indicates the probability of the relation r's existence, N_T and P_T (Figure 2.3) representi the positive and negative classes of an entity pair and TH stands for the introduced threshold class [62].

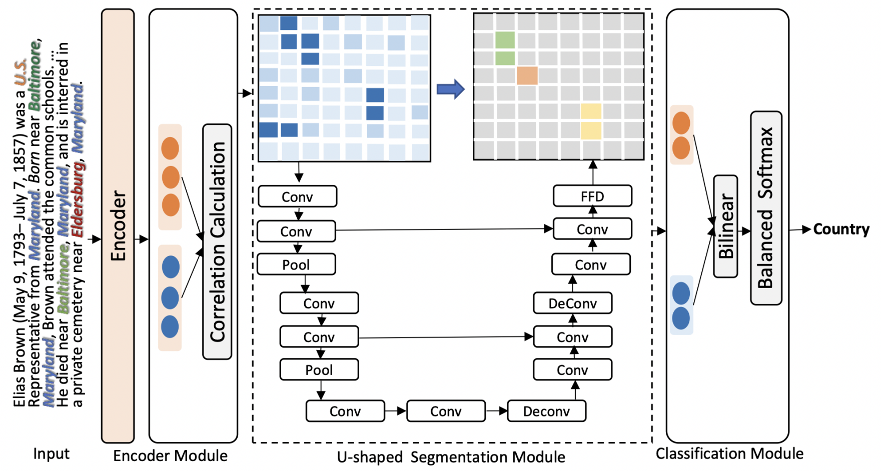

Elias Brown (May 9, 1793– July 7, 1857) was a U.S. Representative from Maryland. Born near Baltimore, Maryland, Brown attended the common schools. ... He died near Baltimore, Maryland, and is interred in a private cemetery near Eldersburg, Maryland.
Input
Encoder
Encoder Module
Correlation Calculation
Conv
Conv
Pool
Conv
Conv
Pool
Conv
Conv
Deconv
Conv
Conv
DeConv
Conv
Conv
FFD
U-shaped Segmentation Module
Bilinear
Balanced Softmax
Country
Classification Module

Figure 2.4: DocuNet Structure[61]

2. **DocuNet**[61]:
Zhang et al. deal with the relation extraction task through a semantic segmentation approach, their model name standing for "Document U-shaped Network" [61]. The DocuNet model has three main steps: An encoder module for gathering the context embeddings, the U-shaped segmentation module and the Classification module [61].

The encoder module has the following structure: After the contextual embeddings H are extracted through a pre-trained encoder module, the entity embeddings e_i $(i \in \{s, o\})$ are obtained through pooling. Afterwards, the entity-level relation matrix $F(e_s, e_o)$ (Figure 2.5) is obtained through the entity embeddings, which marks the relations between the entities in the document on a matrix. The authors use two different approaches for calculating the matrix, using similarity(Eq.(2.12)) and using context(Eq.(2.13)) [61].

$$F(e_s, e_o) = [e_s \odot e_o; \cos(e_s, e_o); e_s W_1 e_o] \tag{2.12}$$

$$F(e_s, e_o) = W_2 H a^{(s,o)} \tag{2.13}$$

Hereby W_1 and W_2 represent weight parameters and $a^{(s,o)}$ stands for the attention weight. In the similarity-based approach, the matrix is calculated through a concatenation of element-wise similarity, cosine similarity and bilinear similarity [61]. The context-based approach uses attention for the calculation [61].

The U-shaped Segmentation Network is comprised of several up- and down-sampling blocks, as seen in Figure 2.4. The output of the module is given with the following equation, with the weight parameters W_3 [61]:

$$Y = U(W_3 F) \tag{2.14}$$

Lastly, the classification module calculates the probability $P(r|e_s, e_o)$ of a relation for an entity pair with the sigmoid activation[61]:

$$P(r|e_s, e_o) = \sigma(z_s W_r z_o + b_r) \tag{2.15}$$

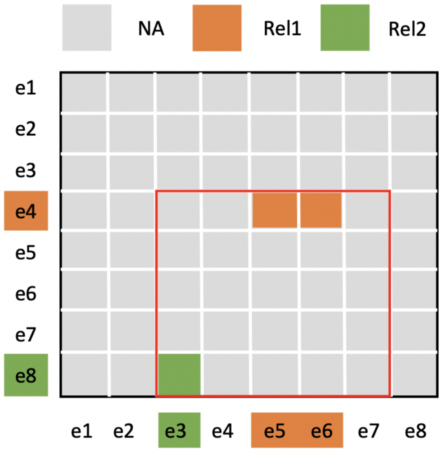


Figure 2.5: Entity-level Relation Matrix for DocuNet[61]

$$z_i = \tanh W_i e_i + Y_{s,o} \tag{2.16}$$

where $i \in \{s, o\}$, where $Y_{s,o}$ represents the entity-pair in the entity-level relation matrix from the output of the U-shaped segmentation network, b_r representing the bias, W_s and W_o standing for the weight parameters for the entity pair. The model uses a variation of circle loss [50] as the loss function[61].

3. **Seq2Rel**[13]:
   Seq2rel is a document-level method for extracting the entities and relations jointly, while applying a sequence-to-sequence technique [13].

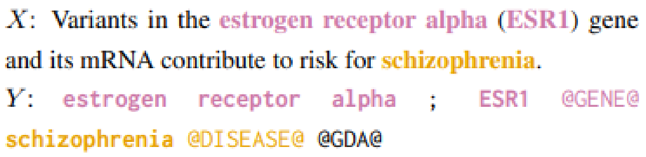
   Their method can be summarized into 3 steps: Linearization, encoding, and the decoding. The first step, linearization, is used to convert the document into a specific format to express the existing relations [13].

   As seen in Figure 2.6, Y, the corresponding sentence for the input X, defines the entities in the sentence X. The elements of the entity span are put together, separated by a semicolon, and different entities are separated by a terminating special token(@...@) that defines their type, a relation ends with a token showing the relation type [13].

   After each contextual embedding is obtained through the encoder, a decoder creates an output with the predictions of entities and relations [13]. The conditional probability is calculated with:

$$p(Y|X) = \prod_{t=1}^{T} p(y_t|X, y_{<t}) \tag{2.17}$$

   where Y represents the target text, X stands for the original text, y_t standing for the i'th token of the text [13].



X: Variants in the estrogen receptor alpha (ESR1) gene and its mRNA contribute to risk for schizophrenia.

Y: `estrogen receptor alpha ; ESR1 @GENE@ schizophrenia @DISEASE@ @GDA@`

Figure 2.6: Seq2rel Linearization[53][13]

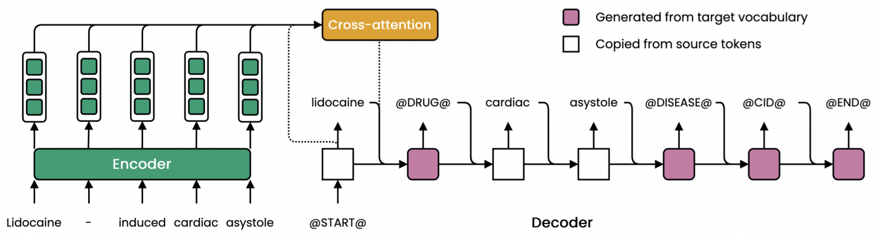


Figure 2.7: Seq2rel Representation[13]

The models loss for training is the following sequence cross-entropy loss, where θ stands for the weight parameters [13]:

$$l(\theta) = -\sum_{t=1}^{T} \log p(y_t|X, y_{<t}; \theta) \tag{2.18}$$

4. **SSAN**[54]:
SSAN model has its foundations on the self-attention concept. The model uses a different entity representation than other models we have represented. The authors create a scheme for the dependencies of the entity mentions with the classes intra+coref, intra+relate, inter+coref, inter+relate, which classify whether the entities are in the same sentence or not (inter or intra) and whether the entities are referencing each other or not(coref or relate) [54].

For predictions, the model extracts the contextual embeddings of the entity pairs considering the aforementioned entity structure proposed by the au-

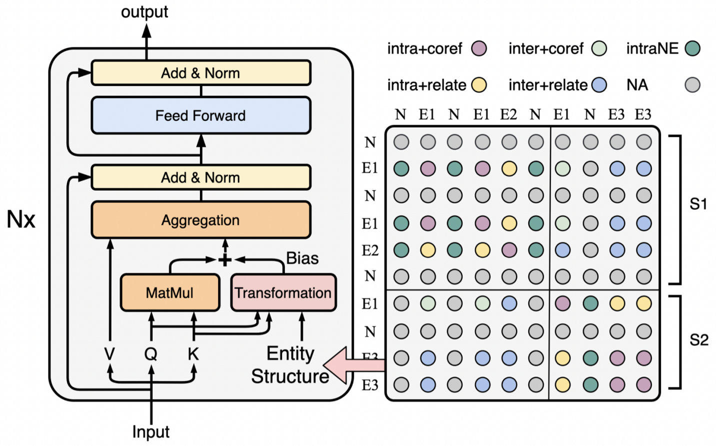


Figure 2.8: SSAN architecture[54]

thors. After pooling is performed on the entity representation, the probability $P_r(e_s, e_o)$ of a relation is calculated through their proposed architecture(Figure 2.8) as follows [54]:

$$P_r(e_s, e_o) = sigmoid(e_s W_r e_o) \tag{2.19}$$

where the e_i for $i \in s, o$represent

The model uses the following cross-entropy loss for training:

$$L = \sum_{<s,o>} \sum_{r} CrossEntropy(P_r(e_s, e_o), \overline{y}_r(e_s, e_o)) \tag{2.20}$$

where the $\overline{y}(e_s, e_o)$ stands for the target label for the entity pair [54].

2.3 Impact Evaluations and Causality

Before diving into the our framework, we should introduce what an impact evaluation is and what importance it has for our research. Evaluations are assessments made on projects, policies or programs specifically for answering questions on the design, implementation or results [12]. According to Gertler et al. [12], an impact evaluation is a specific kind of evaluation that focuses on answering the following: ”What is the impact [...] of a program on an outcome of interest?” [12]. The IE studies are done by several organizations such as International Initiative of Impact Evaluation [24], who also trying to ensure the qualities of the studies.

Through the documentation reports of these interventions, it is possible to infer which results should be expected as an effect of the intervention. Impact evaluations allow the policy makers to be informed on many aspects regarding the future adjustments [12].

IE has a significant importance hereby, as through informing the policy makers, they can make their decisions based on the programs evaluated. Regarding the correlation between causality extraction and the impact evaluations, through causal inference, by assessing the decisions made by previous programs and inferring their results, information can be presented in a compact way to the decision makers. This would also reduce the time that is required to gather the causality manually.

Related Works

Lastly, some works with similar goals should be named before going into the framework. In the only work dealing with interventions we could find, Min et al. create a exclusive corpus for interventions and try extracting the interventions using a CNN model [33]. Their dataset for interventions, however, as mentioned in the first chapter, is not publicly available. Their approach is solely focusing on intervention extraction, and does not deal with the extraction of their effects.

There are also several works that try to incorporate causality extraction into document-level. Mueller and Huettemann propose CauseMiner [37], a work that takes a rule-based approach to extract causal relations using ontology [37]. In their work, causality is extracted using causal patterns after the text is processed into smaller cause and effect sections. The main difference of their work compared to ours is the main task of hypothesis extraction.

In another work, DeepCause [36], the authors build on the results of the CauseMiner and use sequence labeling to extract causality for ontology learning [36]. They use sequence labeling to extract causality in one task, compared to our framework using two tasks NER and RE. Their approach incorporates Part-of-Speech Tags, Bi-LSTMs and CRFs for the task[36].

Another method, CauseNet[18], uses a knowledge base based approach, where a causality graph is created exclusively for causality. Their approach also uses a Bi-LSTM-CRF for the task, which also makes use of linguistic features such as the Part-of-Speech tags [18].

One last method similar to our task is the World Modelers [52], [6], which has a similar methodology to our task, where also a pipeline from different methods is built, on a different domain however.

The lastly mentioned four works have different domains as their focus, which does not fit into our topic, so we are not using the methods.

3. Methodology

In this section, we will be explaining our approach at extracting causality in the domain of impact evaluation. Our proposed approach is consisting of three major steps: Data Collection, Data Labeling (Annotation) and creating a causal extraction pipeline. And our pipeline for causal extraction has five steps: Preprocessing, Tokenization, Named Entity Recognition, Processing for RE and Relation Extraction.

3.1 Data Collection

As the first step of our approach, we gather data that we will be using for our experiment. Following our research goals, we need two datasets: We look for a dataset created for a general domain, and we need a dataset for our specific domain of impact evaluations.

3.1.1 CREST[21][42]

The collection of CREST [21] is one of the largest corpus of collections of causal relation datasets that we are aware of. CREST contains 41,165 samples from 13 data collections, 33,174 of samples being publicly available to us [21]. The CREST collection contains for each sample sentence several columns, of which our interest lies in the following: ”original_id” for the sample number, ”span1” for first entity, ”span2” for second entity, ”label” for causal meaning (0 for no causality, 1 for causality, 2 for counterfactual relations), ”direction” for the direction of the relation (0 for span1 to span2, 1 for span2 to span1), ”context” for the document or the sentence the spans are given, and finally ”idx” for the positions of spans in the sentence [21]. It has to be noted that the CREST dataset has many collections which were originally made for sentence-level causality extraction. Most of the collections in the CREST dataset contain spans that are made of shorter tokens, which are proper nouns, verbs, or sentences in many cases. The datasets do not contain any specific intervention names or exclusive samples for the domain of impact evaluation.

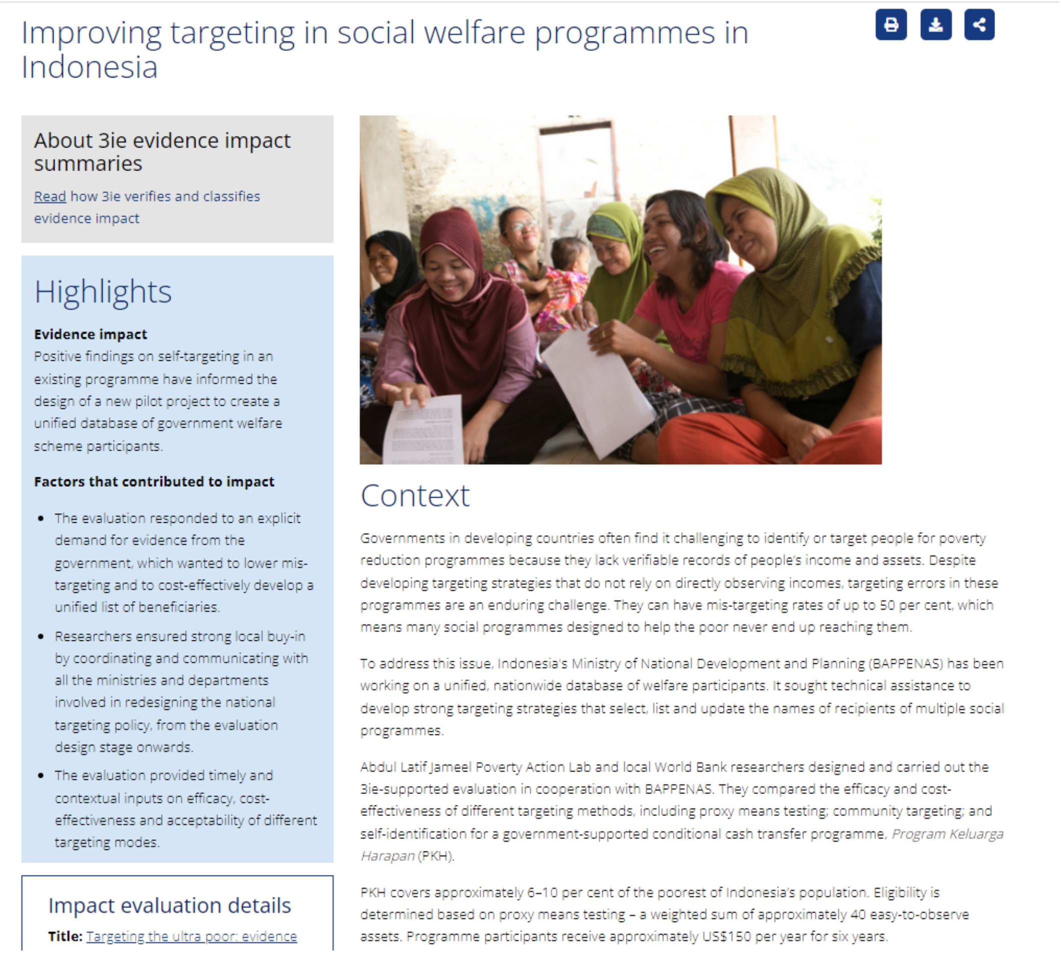
CREST Collections					
ID	Collection	Total Samples	Causal Samples	Non-causal Samples	Year
1	SemEval 2007 Task 4[14]	1,529	114	1,415	2007
2	SemEval 2010 Task 8[19]	10,717	1,331	9,386	2010
3	EventCausality[9]	583	583	-	2007
4	Causal-TimeBank[34]	318	318	-	2007
5	EventStoryLine v1.5[5]	2,608	2,608	-	2016
6	CaTeRS[35]	2,502	308	2,194	2016
7	BECauSE v2.1[10]	729	554	175	2017
8	Choice of Plausible Alternatives[47]	2,000	1,000	1,000	2011
9	Penn Discourse Treebank 3.0[43]	7,991	7,991	-	2019
10	BioCause[32]	844	844	-	2013
11	Temporal and Causal Reasoning[41]	172	172	-	2018
12	Adverse Drug Effects[16]	5,671	5,671	-	2012
13	SemEval 2020 Task 5[57]	5,501	5,501	-	2020

Table 3.1: Datasets in CREST collection[21][42]
(The collection Penn Discourse Treebank 3.0 is not publicly available)

3.1.2 Impact Evaluation Dataset

For our domain specific dataset creation, we gather summaries of several impact evaluations done by International Initiative for Impact Evaluation(3ie) published on 3ieimpact.org [24], [11]. We use all 60 publicly available impact evaluation summaries to create our dataset. The summaries present an evaluation of an intervention. The documents are structured like the following [25]:

1. Title : Name of the article
2. Highlight section: Very short summary of the evaluation and effects
    (a) Evidence impact: What has the evaluation found about the intervention, effects
    (b) Factors contributed to impact: Factors that helped the evaluation process
3. Impact evaluation details: Basic information such as title, interventions name and authors
4. Context: First, background information is presented. Then the intervention is explained in detail.



Improving targeting in social welfare programmes in Indonesia

About 3ie evidence impact summaries

Read how 3ie verifies and classifies evidence impact

Highlights

Evidence impact
Positive findings on self-targeting in an existing programme have informed the design of a new pilot project to create a unified database of government welfare scheme participants.

Factors that contributed to impact

- The evaluation responded to an explicit demand for evidence from the government, which wanted to lower mis-targeting and to cost-effectively develop a unified list of beneficiaries.
- Researchers ensured strong local buy-in by coordinating and communicating with all the ministries and departments involved in redesigning the national targeting policy, from the evaluation design stage onwards.
- The evaluation provided timely and contextual inputs on efficacy, cost-effectiveness and acceptability of different targeting modes.

Impact evaluation details

Title: Targeting the ultra poor: evidence

Context

Governments in developing countries often find it challenging to identify or target people for poverty reduction programmes because they lack verifiable records of people's income and assets. Despite developing targeting strategies that do not rely on directly observing incomes, targeting errors in these programmes are an enduring challenge. They can have mis-targeting rates of up to 50 per cent, which means many social programmes designed to help the poor never end up reaching them.

To address this issue, Indonesia's Ministry of National Development and Planning (BAPPENAS) has been working on a unified, nationwide database of welfare participants. It sought technical assistance to develop strong targeting strategies that select, list and update the names of recipients of multiple social programmes.

Abdul Latif Jameel Poverty Action Lab and local World Bank researchers designed and carried out the 3ie-supported evaluation in cooperation with BAPPENAS. They compared the efficacy and cost-effectiveness of different targeting methods, including proxy means testing; community targeting; and self-identification for a government-supported conditional cash transfer programme, *Program Keluarga Harapan* (PKH).

PKH covers approximately 6–10 per cent of the poorest of Indonesia's population. Eligibility is determined based on proxy means testing – a weighted sum of approximately 40 easy-to-observe assets. Programme participants receive approximately US$150 per year for six years.

Figure 3.1: Example impact evaluation summary [25]

5. Evidence: Explains the effects of the intervention found by the evaluation.
6. Evidence impact: Presents effects of the evaluation, such as law changes, reconsiderations of the intervention, continuation of the intervention.
7. Suggested citation and references: Citations and references

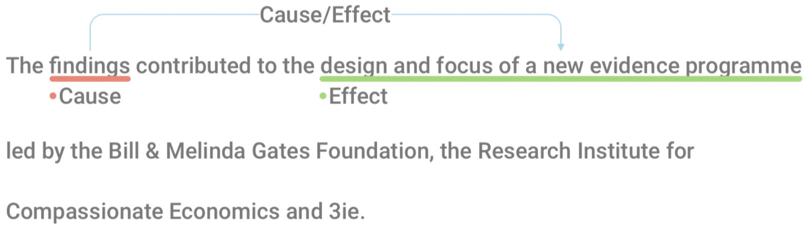
It is worth noting that the intervention names used in the documents are in some cases made of words from other languages of the interventions origin country, and their abbrevations of the interventions are used in many places in the evaluation reports. In other cases they can be domain-specific word groups like ”monetary incentives” or ”food vouchers” [11].

After gathering the data, we move on to the next step of data labeling.

3.2 Data Annotation

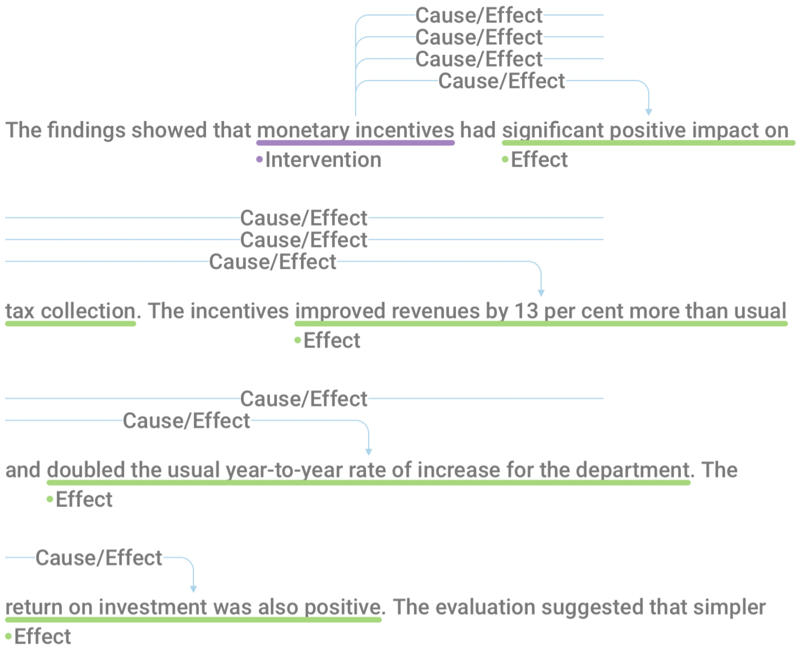
For the evaluation of the model, we need to annotate the documents we have collected. For this process we use Doccano [40], an annotation tool for many tasks like sequence labeling and NER. Our labeling scheme consists of 3 span types and 1 relation type:

1. Cause (Span)
2. Effect (Span)
3. Intervention (Span)
4. Cause/Effect (Relation)



The findings contributed to the design and focus of a new evidence programme led by the Bill & Melinda Gates Foundation, the Research Institute for Compassionate Economics and 3ie.

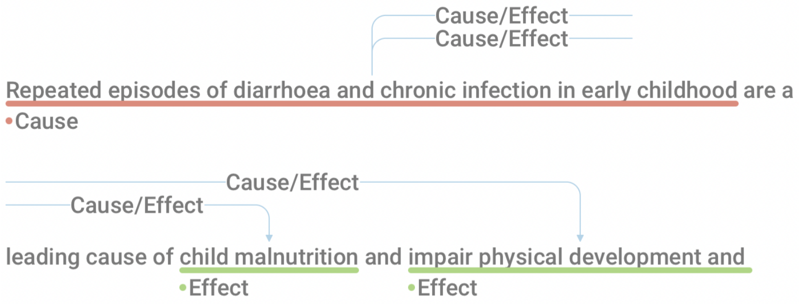
Figure 3.2: Example Annotation (Cause Label-1)
Cause annotation of ”findings” represents the evaluation paper and is domain related.



The findings showed that monetary incentives had significant positive impact on tax collection. The incentives improved revenues by 13 per cent more than usual and doubled the usual year-to-year rate of increase for the department. The return on investment was also positive. The evaluation suggested that simpler

Figure 3.3: Example Annotation (Intervention Label)
The intervention is the ”monetary incentives” has the effects seen above.

We use the label ”Intervention” for labeling different mentions of the intervention. We use the ”Cause” label for two types of spans: Spans containing any general causal meaning and spans mentioning the study itself. The label ”Effect” is for any kind of effect of a causal relation, both domain-specific and general, also as an effect to both ”Intervention” and to ”Cause”. Lastly, we create one relation type called ”Cause/Effect” that we use to annotate the existing causal relations between a causal pair. This approach with a relatively low amount of labels for spans allows us to keep the labeling scheme simple. Example annotations can be seen in Figures 3.2,



Cause/Effect
Cause/Effect

Repeated episodes of diarrhoea and chronic infection in early childhood are a
•Cause

Cause/Effect
Cause/Effect

leading cause of child malnutrition and impair physical development and
•Effect •Effect

Figure 3.4: Example Annotation (Cause Label-2)
Cause and effect annotations in this example are from a general domain that is closely related to the topic, so it is also included.

3.3, 3.4 (from Doccano [40]). In most cases, the effects that we have annotated are more likely to include longer spans with verbs, while the causes we have annotated are shorter in comparison in token length.

It has to be noted that our labeling scheme is very different from general causality datasets like CREST. Our dataset contains many spans consisting of longer sequences of tokens that are not a complete sentence, whereas a big part of the CREST dataset contains spans consisting of one or two words, proper nouns and verbs in most cases, and several collections contain supporting sentences.

After the annotation process, we will be using the created annotated dataset in our Causality Extraction Pipeline, for training and also for predictions.

3.3 Causality Extraction Pipeline

This section presents the main pipeline we propose for our causality extraction model. The pipeline consists of 5 main steps: Preprocessing, Tokenization, Named Entity Recognition, Processing for the Relation Extraction and finally Relation Extraction.

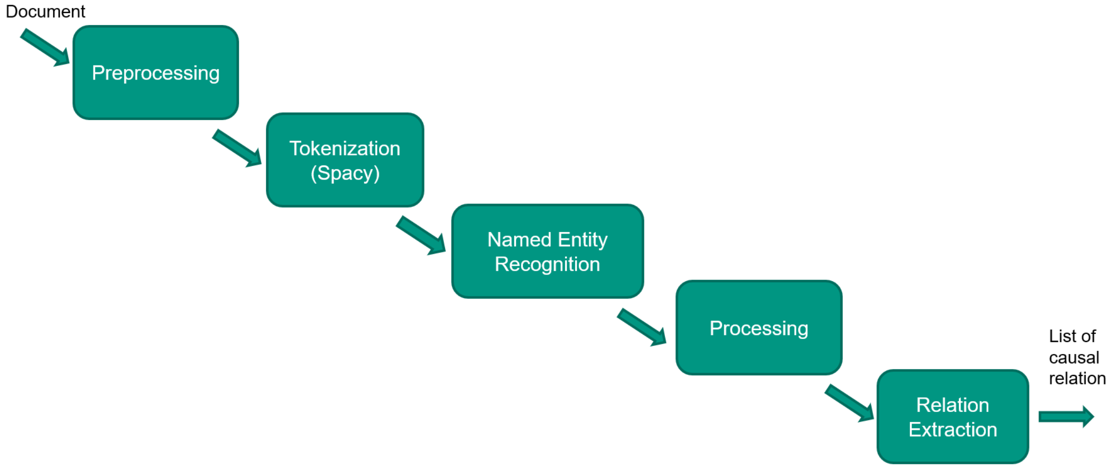


Figure 3.5: Representation of Proposed Pipeline

3.3.1 Preprocessing

The first step of our causality extraction pipeline is the preprocessing. This step aims to shape the dataset we have created to fit it into the named entity recognition model.

During the preprocessing, we split the summaries into several sections: Title, Highlights and Details, Context, Evidence, Evidence Impacts, Suggested citation. This split, while not damaging any labeling structure created in the data labeling section (there are no relations annotated going across different paragraphs), allows leaving out unnecessary parts of the text being fed into the causality extraction pipeline, such as the ”Title” and the ”Suggested citation and references” parts.

Then the dataset is processed into a format of a Document-level Relation Extraction dataset, DocRED [59], for a quicker matching during the transition between the Named Entity Recognition and the Relation Extraction. More information on the DocRED dataset will be presented in the corresponding section: 3.3.4.

One problem to consider is that the use of an intervention label during the annotation creates a situation where if the CREST dataset is used for training in any of the models, then the model will not be able to predict any specific intervention. CREST dataset does not include any labels other than ”span1” and ”span2” for cause and effect. Following this, all labels of ”Intervention” are changed to ”Cause” during this step for a better comparison between the two datasets.

3.3.2 Tokenization

The models that we are going to use tokens as their inputs. A token specifies the smallest piece of information that can be described as a word in our case. In order to continue with the Named Entity Recognition model, we have to tokenize the sentences of the documents.

Tokenization is a process of creating tokens with a basic set of rules. There are several libraries that can be used in Python for tokenization. We use spaCy[20] for our purposes. The paragraphs are first separated into sentences, and then into its tokens using spaCy’s internal modules.

3.3.3 Named Entity Recognition

In this step, for the Named Entity Recognition, we are taking a different approach than the definition in Section 2.2.1: Our goal in the Named Entity Recognition is going to be extracting the cause entity candidates and effect entity candidates, by training the model with the cause and effect spans of the datasets CREST or the Impact Evaluations separately. The main reason for this is the difference between our labeling scheme and the general NER approaches. As we have mentioned before, most of the collections in the CREST dataset contain spans that are made of shorter tokens, compared to the Impact Evaluation dataset. The existing NER models extract entities that are proper nouns, whereas our labeled data consists of entities of longer spans, thus making the normal approach of extracting main entities more difficult. This can be seen in Figures 3.2, 3.3 and 3.4. Considering these problems, we are using a Named Entity Recognition that is not focusing on entities with smaller and more compact labels like spaCy NER, instead we are using *Sequence Labeling(Tagging)*. To perform this, we are using an implementation [26] of the Bi-LSTM-CRF model mentioned in 2.2.1.

To obtain an acceptable input, we create a new text file containing every tokenized sentence of each document and a list of tags for each token. We use the tags 'C' for cause, 'E' for effect and 'O' for other unrelated entities. The list of tags for cause and effect are to be used for the evaluation, for the prediction only the tokenized sentence will be used. The model has creates its predictions as a list of predicted tags. Token sequences corresponding to a continuous sequence of the same tags of 'Cause' and 'Effect' will be taken as complete entities. The entity positions and names are then used to get the document ready for the Relation Extraction.

3.3.4 Processing for Relation Extraction

The entity candidates for cause and effect we get from the Named Entity Recognition model must first be processed before being put into the relation extraction models for the relation extraction. All of the models that we have considered for our approach have an accepted dataset as their input in common: DocRED [59]. DocRED is a dataset created specifically for the domain of Relation Extraction. We explain the general structure of DocRED in the following paragraph based on the work [59]:

The dataset is a collection of documents parsed into JSON file where four keys exist for each document: "title", "sents", "vertexSet" and "labels". "title" stands for the title of the document, it is a string. "sents" is a nested list of strings, it is a list of all sentences in a document, where a sentence(inner list) is also a list of every token in that sentence. "vertexSet" consists of lists of entities, where each mention of one entity is grouped inside the same list. "labels" is a list that contains the valid relations between the entities inside the "vertexSet". Each entity inside the "vertexSet" has the keys of "name", "sent_id", "pos", and "type", where each represent in order the entity span, the sentence number inside the "sents" list where the entity span appears (short for sentence ID), a list of start and end indexes of tokens of the entity span in the sentence, and a string for the type of the entity. The relations inside "labels" also contain 4 keys: "h", "t", "r" and "evidence". "h" stands for "head" and gives the index of the starting entity of the relation from the "vertexSet". "t" stands for "tail" and represents the index of the ending entity inside the "vertexSet". "r" specifies the type of the relation. Lastly "evidence" is a list for the indexes of the sentence ID's for the head and tail entities, it is a list for the information required to deduce the relation. During the processing, considering many mentions of interventions can appear in the text, they are also grouped together as the same entity in the list of entities("vertexSet").

In case of CREST, there are many samples where the sample only contains one sentence, as the datasets like SemEval 2007 Task 4 [14] were made for sentence-level relation extraction exclusively. On these occasions, we have only one sentence list inside the "sents" list for one document. The CREST dataset also contains the span position information for each sample, so we convert these information to token positions. After the predictions of the NER model are parsed into the DocRED format, they are given to the Relation Extraction models.

3.3.5 Relation Extraction

As the last part of our pipeline, the relation extraction methods aim to predict if the given entity candidates for cause and effect have a causal relation between them, with its direction. As a result, the models will be giving a list of relations that it has predicted. One important feature of the document-level relation extraction methods are the special focus on the mentions of entities. As one single entity can be mentioned in a document in different locations with different language, the separate entity mentions must be found and regarded as part of an entity. We try to incorporate the models mentioned in Section 2.2.2. The model configurations are to be changed to fit into our task, where the only relation is going to be the Cause/Effect relation, and the only subject and object entities (e_s, e_o) will be Cause and Effect entities respectively.

4. Evaluation

In this chapter, we are presenting our evaluation metrics, experimental setup and the results of our experiment.

4.1 Evaluation Metrics

There are several commonly used metrics to evaluate a machine learning algorithm. Before going into the evaluation of the methods, we should be inspecting which metrics the methods we have used are evaluated with. Metrics that we will be looking at in this section are: precision, the F1-score and recall.

Confusion Matrix The confusion matrix is a classification scheme for the instances. During the evaluation we categorize the instances(example sentences) that the methods evaluate on the data sets as follows:

The classes True Positive, False Positive, False Negative and True Negative are used for demonstrating the correctness of the prediction for a sample. True Positive (tp) gives out the correctly predicted samples that are actually positive samples, False Negative stands for the positive samples that were classified as negative instances (fn)[56]. If the class is actually a sample that is negative, then this instance is classified as a False Positive instance(fp), and lastly in the case where the instance is classified as negative when it actually is a negative sample, the instance is a True Negative (tn)[56] (see Table 1). The metrics can be derived from the confusion matrix.

Table 4.1: Confusion Matrix

		Actual Instance	
		positive	negative
Predicted Instance	positive	True Positive(tp)	False Positive(fp)
	negative	False Negative(fn)	True Negative(tn)

1. Precision: Precision is the rate of correctly predicted positive instances among all of the actual positive instances [56].

$$precision = \frac{tp}{tp + fp} \tag{4.1}$$

2. Recall: Recall is the rate of the correctly predicted positive samples among all samples that were predicted as positive[56].

$$recall = \frac{tp}{tp + fn} \tag{4.2}$$

3. F-Score: The F-score is a metric that is calculated through precision and recall. The F_1-score is the most commonly used variation of the score among most works. According to Sasaki, F_1-Score can be defined as a harmonic mean between precision (P) and recall (R) [48], as in formula (4.4):

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \tag{4.3}$$

4.2 Experimental Setup

In this section our preparation for the experiments will be presented.

4.2.1 Datasets for training

This subsection will be presenting the changes we have made in the datasets during the evaluation phase. The amount of samples in each datasets can be found in Table 4.2.

4.2.1.1 Impact Evaluation Dataset(IE)

With our objectives in mind, we aim to see how differently the models we have found can infer the causal relations between the interventions and their effects when they are trained on a dataset on the special field of impact evaluations. For this purpose we will be using the dataset we have labeled from chapter 3 as both training and test sets.

As mentioned before, the IE dataset is consisting of sections of documents as samples. Each document has around 4 or 5 sections, in total making around 237 samples with unrelated sections like titles left out.

	CREST-1	CREST-2	IE (full)	IE Training Set	IE Test Set
Samples	20,613	31,540	237	201	36
Total Valid Relations	20,613	31,540	953	796	157
Paper Count	-	-	60	51	9

Table 4.2: Dataset Statistics

First, we split the IE dataset into training and test sets with an 85%/15% partition. The training set contains 51 papers with 201 samples, and the test set contains 9 documents with 36 samples in total. We will use this test set with 9 documents as our test set for every evaluation step.

For the NER model, we parse the sets into accepted .txt formats containing the sentence and tags lists. The NER model also requires a vocabulary file containing all vocabulary of the training set, we parse it from the training set. For the RE models, we parse the sets into the DocRED format. The training set with 51 papers is then split into 5 sets to perform 5-fold cross-validation.

To fit the data in all of the models, we limit the maximum sequence (token) length of all documents at 1024.

4.2.1.2 CREST[21]

We use two variations of the CREST dataset for the experiment. For both of the variations, we first remove unusable samples containing ”” or ””” (empty string) in any of its spans. We use these variations only for training and validation of the models.

In the first variation CREST-1 we omit a considerable amount of samples out for our experiment due to tokenization problems. The CREST-1 set contains in total 20613 rows, where each row contains exactly one relation. We also remove the medical datasets that we consider containing too detailed and domain-specific medical samples: The BioCause corpus [32] with 844 samples and the ADE(Adverse Drug Effects) corpus[16] with 5671 samples.

In the second variation CREST-2, we perform the tokenization, but the sentences are not separated: the samples contain all of the text as one sentence. In this case, the variation contains 31540 rows, where the rows also contain exactly one relation. We

For the NER model, we parse the two variations into the accepted .txt formats, and their vocabulary files are created.

For the RE models, the we parse the sets into the DocRED format. CREST-1 dataset is segmented into its sentences, the ’sents’ column is made of multiple sentences if the text is made of several sentences. In the CREST-2 variation the text is parsed as a single sentence inside the ’sents’ column. We then split both variations into 5 subsets for the 5-fold cross-validation.

The same way as we have done in IE dataset, we limit the maximum sequence length of all documents to 1024.

	CREST-1		CREST-2		IE
Models	ATLOP	DocuNet	ATLOP	DocuNet	ATLOP&DocuNet
Batch Size	32	32	16	16	4
Learning Rate	5x10^5	5x10^5	5x10^5	5x10^5	5x10^5
Epochs	10	10	10	10	20

Table 4.3: Hyperparameter Settings

4.2.2 Experimental Settings and Procedure

The implementations are mostly using PyTorch. To do our experiment, we are using the environment of bwUniCluster2.0(The authors acknowledge support by the state of Baden-Württemberg through bwHPC). The environment we use has 1 GPU, 10 CPUs and 90 GB RAM.

The Bi-LSTM-CRF model is trained with 1024 maximum sequence length, with epoch numbers of 30, 25 and 70 respectively for CREST-1, CREST-2 and IE datasets. For the evaluation of the Bi-LSTM-CRF model, we are calculating the scores based on single predictions of each token, if a token is inside a cause or effect span, and it is predicted by the model as such, we classify this as a correct prediction. The metrics are calculated separately for Cause and for Effect tags (C and E).

The RE models are configured to fit our data. For both of the models, we use two relation types: Cause/Effect(”P9999”) and Na(”Na”). The entity types for the experiment are ”CAUSE” and ”EFFECT”. The model hyperparameters can be seen in Table 4.3. We use BERT[8] as the pretrained model for both ATLOP and DocuNet. 5-fold cross-validation is performed during the training of the RE models. For the 5-fold cross-validation, we are performing the following procedure: First, we separate the training sets into five and do five training iterations, where we use one of the separated groups as the validation set in each iteration. The IE test set is the test set for all of our models, we run the models for the prediction of the IE test set, and the mean averages of the scores obtained from this prediction are calculated as the final results. An error interval is also calculated through standard deviation.

4.3 Problems

Before presenting our results, we introduce the problems we have come across before evaluating our models.

The methods seq2rel and SSAN were found to be unsuitable to our task. The model seq2rel had the scores of 0% precision, 0% recall and 0% F_1-score. The SSAN model was unable to calculate the training loss for the CREST dataset, and gave the value ”NaN” during the training. Because of these reasons, we are not including these models in the following results of the evaluation.

Another problem has occurred during the tokenization of the documents of the CREST-1 dataset as mentioned in the previous section. Due to tokenization problems, the collection SemEval 2020 Task 5 [57] was completely disregarded for the

Dataset-Entity	F1	Precision	Recall
CREST1-Cause	5%	8%	4%
CREST1-Effect	6%	29%	3%
CREST2-Cause	5%	8%	4%
CREST2-Effect	6%	26%	3%
IE-Cause	13%	40%	7%
IE-Effect	28%	42%	22%

Table 4.4: NER Results Table

CREST-1 iteration, several other samples with problematic tokenization were also removed. Due to time constraints, this problem could not be resolved.

Also, during the evaluation, some problems occurred that must be mentioned. The amount of epochs were held low, as the results were found not to be improving after the given amounts of epochs. Due to ATLOP's missing log-files, we cannot show any loss graphs for the ATLOP model. As for DocuNet, we will be presenting a table for the changes in training and validation loss for one cross-validation set. Due to time restrictions, we cannot present the losses for the other cross-validation sets, however, the patterns in every cross-validation set appears to be similar. Due to the batch size selection and the implementation of the result log file of DocuNet, with the losses of IE datasets only several loss values were printed, due to time restrictions the rest could not be implemented by us.

One last thing to mention is that the change of 'Intervention' labels to 'Cause' labels after the data annotation step has also caused us to not be able to distingush general causal entities in the IE dataset, we will be regarding them as both causes.

4.4 Model Performances

During our experiments, we evaluate the NER and RE sections separately, we test the RE models by giving the models the annotated entities with the documents.

4.4.1 Named Entity Recognition

The results for the evaluation of the Bi-LSTM-CRF can be found in Table 4.4 and Figure 4.1. The calculation for Effect and Cause labels are done separately and we are inspecting the predictions on the token level.

With a first look on the graphs, we can see that on average, the IE dataset has resulted in a better overall performance on both Cause and Effect labels. The CREST-1 and CREST-2 datasets have all of their values below 30%, and most of the scores are below 10%. On the CREST-1 dataset, it can be seen that the Bi-LSTM-CRF model obtains an F_1 score of 5%, a precision of 8% and a recall of 4% for Cause predictions. On Effect tags, the model obtains the scores of 6% F_1-score, 29% precision and 3% recall. When trained with the CREST-2 dataset, we observe a small decrease in precision on the Effect labels, with the value being 26%. All of the other metrics remain the same.

For the IE dataset, we observe an improvement for each metric on both labels. Trained with the IE set, the model gets a precision score of 40%, a recall score of 7% and an F$_1$-score of 13% for the Cause tags. For the Effect label the following scores are obtained: 42% precision, 22% recall and 28% F$_1$-score.

Several example predictions for each dataset can be found below:

Example Sentence 1: [’Findings’, ’on’, ’the’, ’effectiveness’, ’of’, ’food’, ’vouchers’, ’have’, ’informed’, ’World’, ’Food’, ’Programme’, ’(’, ’WFP’, ’)’, ’interventions’, ’to’, ’improve’, ’nutrition’, ’and’, ’food’, ’security’, ’of’, ’refugee’, ’populations’, ’in’, ’Ecuador’, ’.’]

Correct tokens for sentence 1: [’C’, ’C’, ’C’, ’C’, ’C’, ’C’, ’C’, ’O’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’O’]

Prediction of CREST-1: [’O’, ’O’, ’O’, ’C’, ’O’, ’O’, ’E’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’]

Prediction of CREST-2: [’O’, ’O’, ’O’, ’C’, ’O’, ’O’, ’E’, ’E’, ’E’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’]

Prediction of IE: [’O’, ’O’, ’O’, ’C’, ’C’, ’C’, ’C’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’O’]

In this example, we can see that both CREST variations have resulted in a prediction of a short sequence of tokens, where in case of CREST-2, the model has predicted three effect entities. The CREST-1 dataset was able to predict *effectiveness* as a cause candidate and *vouchers* as an entity candidate, whereas CREST-2 predicts that *vouchers*, *have* and *informed* as candidates. We observe that in case of IE, the predicted labels are much longer and contains more tokens. The prediction of IE annotates each token of the sequences *effectiveness of food vouchers* as cause candidate and *have informed World Food Programme (WFP) interventions to improve nutrition and food security of refugee populations in Ecuador.* as an effect candidate.

Example Sentence 2: [’The’, ’evaluation’, ’found’, ’that’, ’transfers’, ’through’, ’all’, ’modalities’, ’reduced’, ’controlling’, ’behaviour’, ’amongst’, ’men’, ’and’, ’physical’, ’or’, ’sexual’, ’violence’, ’by’, ’6’, ’to’, ’7’, ’percentage’, ’points’, ’.’],

Correct tokens for sentence 2: [’O’, ’O’, ’O’, ’O’, ’C’, ’C’, ’C’, ’C’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’E’, ’O’],

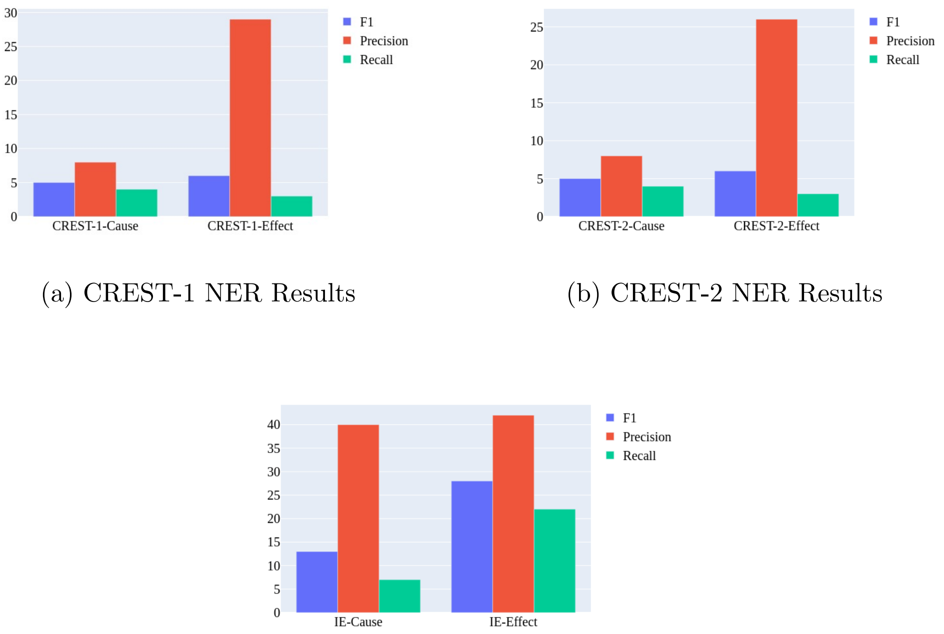
Prediction of CREST-1: [’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’E’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’],

Prediction of CREST-2: [’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’],

Prediction of IE: [’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’, ’O’],

In this example, we see a prediction where the sequence *transfers through all modalities* was originally labeled as an ’Intervention’. Here, we observe that all datasets have resulted in a failed prediction for all cause and effect labels.

Figure 4.2 describes the training and evaluation loss values for the datasets CREST-1, CREST-2 and IE in order. The iterations are equivalent to epoch number times



(a) CREST-1 NER Results

(b) CREST-2 NER Results

(c) IE NER Results

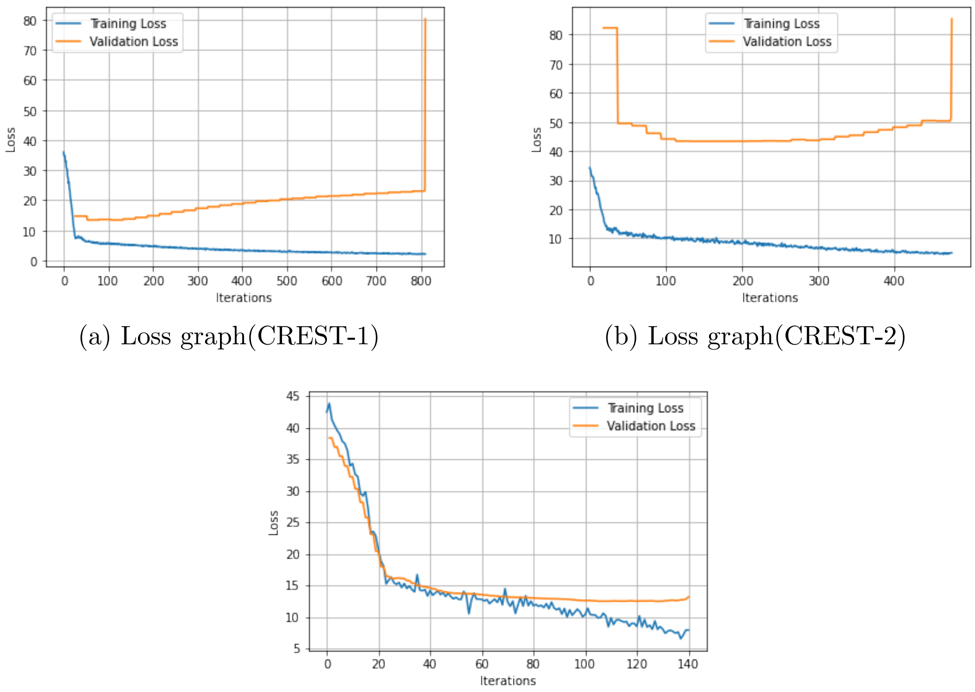
Figure 4.1: NER Results

steps in an epoch, while the final peaks in the graphs represent the final testing step that was done by the model internally with a separate test split from the CREST set. With CREST-1 dataset, we see that the training loss is converging to 0, while the validation loss starts slowly increasing after a short period of decline and then stagnation. We can see the same pattern with the CREST-2 dataset, where the training loss starts around the same level of CREST-1 and converges in time. The validation loss increases after a period of decline and stagnation. With IE, both losses fall in a similar way. After around 60th iteration the training loss continues decreasing, the validation loss decreases very slowly, with small fluctuations happening in the final iterations. From the final testing steps of every graph we observe that the loss obtained from the internal test set is much higher than the final validation losses in case with CREST-1 and CREST-2, it results in a major peak in the final iteration of the graphs. With IE, we observe that the final test loss value is very close to the last validation loss value.

4.4.2 Relation Extraction

The relation extraction model results can be found in Table 4.5 and Figure 4.3. The figures visualize the scores for each dataset, where the bars are presenting the mean average of the 5-fold cross-validation, and the intervals are the error intervals. Looking at the results, it can be seen that the overall scores for the CREST1 and CREST2 datasets are distinctly lower than the scores of IE dataset, except for the recall values.

After training with the CREST-1 dataset, we observe that the predictions have the mean scores of 26.26% F_1-score, 17.56% precision and 53.12% recall for ATLOP. As



(a) Loss graph(CREST-1)

(b) Loss graph(CREST-2)

(c) Loss graph(IE)

Figure 4.2: Loss Graphs for NER

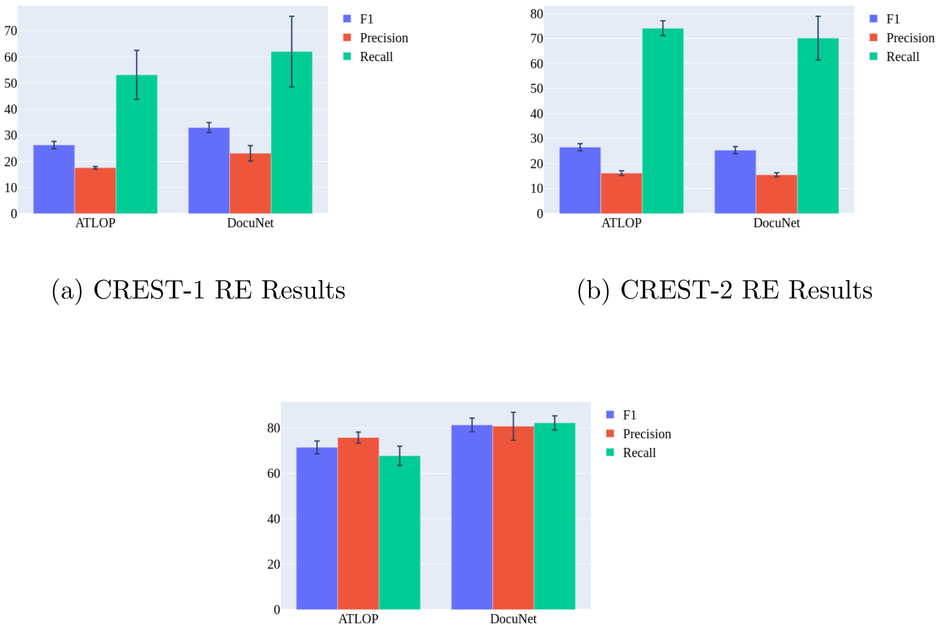
Model	F1	Precision	Recall
CREST-1-ATLOP	26.26 ± 1.38	17.56 ± 0.45	53.12 ± 9.37
CREST-1-DocuNet	32.90 ± 1.93	23.08 ± 2.96	62.04 ± 13.51
CREST-2-ATLOP	26.56 ± 1.45	16.18 ± 0.95	74.14 ± 2.95
CREST-2-DocuNet	25.37 ± 1.43	15.52 ± 0.82	70.19 ± 8.75
IE-ATLOP	71.50 ± 2.84	75.83 ± 2.44	67.77 ± 4.27
IE-DocuNet	81.36 ± 3.00	80.83 ± 6.13	82.29 ± 3.10

Table 4.5: RE Results Table

for DocuNet, we observe 32.90% for F_1-score, 23.08% for precision and 62.04% for recall. With the CREST-2 dataset, the scores for the ATLOP model are 26.56% F_1-score, 16.18% precision and 74.14% recall, while the DocuNet results in 25.37% for F_1-score, 15.52% for precision and 70.19% for recall. The error intervals on F_1-score and precision are relatively lower than the intervals on the recall for every dataset, while we observe that CREST-1 has the largest error intervals among all datasets.

The IE dataset, as seen from the results, has the highest mean averages on every metric except for recall with ATLOP. The IE dataset has the means of 71.50% for F_1-score, 75.83% for precision and 67.77% for recall on ATLOP. As for DocuNet, the scores of 81.36% for F_1-score, 80.83% for precision and 82.29% for recall can be observed.

The changes in losses through the epochs can be seen in Table 4.7 and 4.6. As mentioned before, the losses are available for one set for each dataset. For CREST-1 and CREST-2, the averages of all steps in each epoch are presented. We observe that the training losses on CREST sets start relatively high, they all are decreasing



(a) CREST-1 RE Results

(b) CREST-2 RE Results

(c) IE RE Results

Figure 4.3: RE Results

towards zero. The validation losses, on the other hand, start from values lower than one. We observe that the validation losses increase in the beginning, the loss decreases in both cases in one step only, after which we can observe an upwards trend.

On the other hand, with the IE dataset, the loss begins decreasing in the beginning down to 0.308, after which the loss starts to increase with several decreasing steps in some epochs. We also observe a decline in training loss in the available training data.

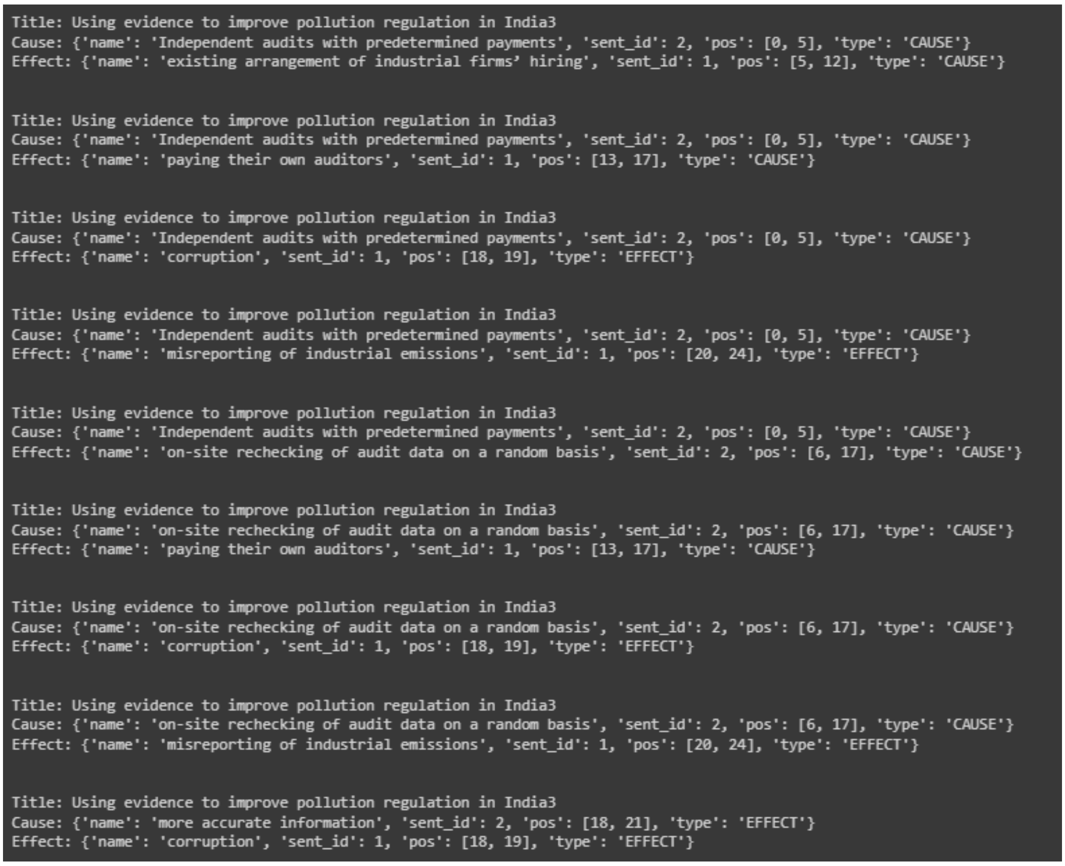
The figures 4.5 and 4.4 visualize several predictions by the ATLOP model using CREST-1 and IE datasets. We observe that with CREST-1, the model predicts a causal relation between two entity candidates of cause.

Dataset-Loss Epochs	CREST-1-Training	CREST-1-Validation	CREST-2-Training	CREST-2-Validation
1	381.146	0.185	457.695	0.289
2	167.919	0.194	256.373	0.385
3	127.108	0.209	179.307	0.345
4	107.644	0.197	124.146	0.523
5	90.634	0.255	75.737	0.561
6	76.682	0.325	55.482	0.830
7	60.075	0.336	37.315	0.913
8	31.937	0.587	23.911	0.986
9	14.235	0.592	16.013	0.961
10	6.619	0.654	10.730	1.071

Table 4.6: DocuNet Losses(Set-1/CREST)

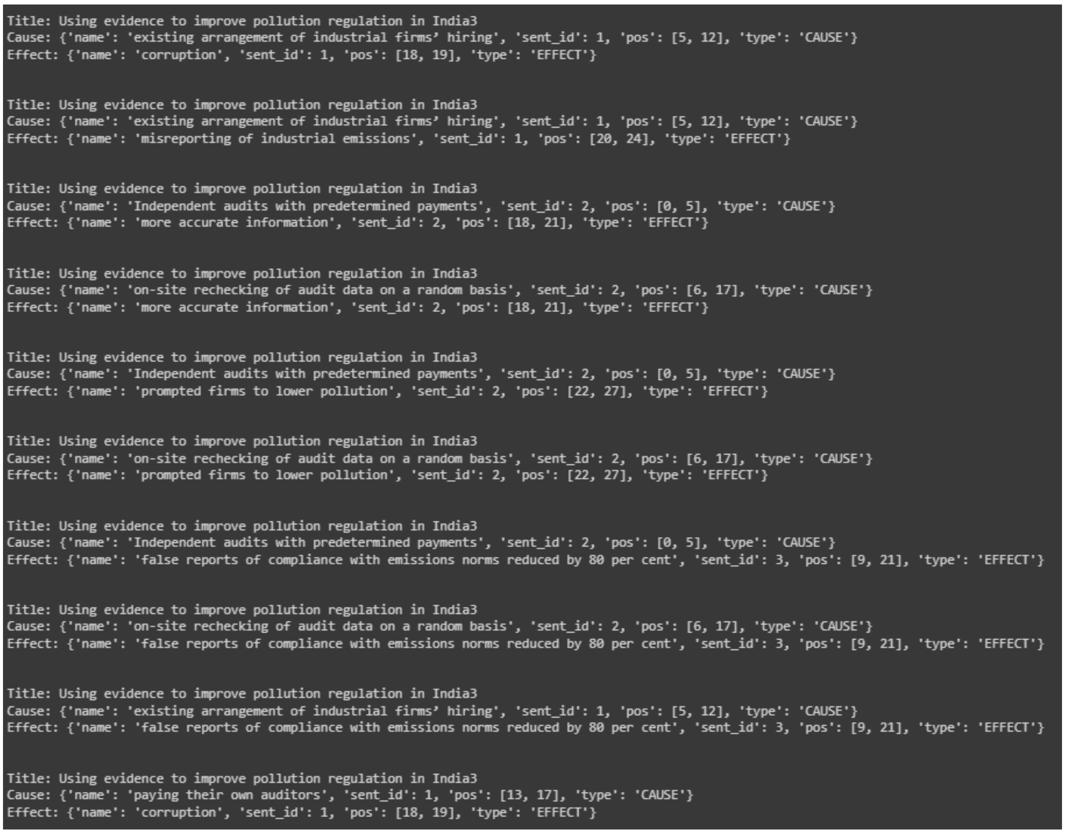
Dataset-Loss Epochs	IE-Training	IE-Validation
1	-	0.708
2	-	0.473
3	614.717	0.427
4	-	0.371
5	204.981	0.308
6	-	0.395
7	-	0.532
8	99.499	0.533
9	-	0.437
10	59.182	0.733
11	-	0.553
12	-	0.645
13	34.693	0.618
14	-	0.594
15	17.225	0.658
16	-	0.753
17	-	0.797
18	10.387	0.781
19	-	0.801
20	6.255	0.785

Table 4.7: DocuNet Losses(Set-3/IE)



```
Title: Using evidence to improve pollution regulation in India3
Cause: {'name': 'Independent audits with predetermined payments', 'sent_id': 2, 'pos': [0, 5], 'type': 'CAUSE'}
Effect: {'name': 'existing arrangement of industrial firms’ hiring', 'sent_id': 1, 'pos': [5, 12], 'type': 'CAUSE'}


Title: Using evidence to improve pollution regulation in India3
Cause: {'name': 'Independent audits with predetermined payments', 'sent_id': 2, 'pos': [0, 5], 'type': 'CAUSE'}
Effect: {'name': 'paying their own auditors', 'sent_id': 1, 'pos': [13, 17], 'type': 'CAUSE'}


Title: Using evidence to improve pollution regulation in India3
Cause: {'name': 'Independent audits with predetermined payments', 'sent_id': 2, 'pos': [0, 5], 'type': 'CAUSE'}
Effect: {'name': 'corruption', 'sent_id': 1, 'pos': [18, 19], 'type': 'EFFECT'}


Title: Using evidence to improve pollution regulation in India3
Cause: {'name': 'Independent audits with predetermined payments', 'sent_id': 2, 'pos': [0, 5], 'type': 'CAUSE'}
Effect: {'name': 'misreporting of industrial emissions', 'sent_id': 1, 'pos': [20, 24], 'type': 'EFFECT'}


Title: Using evidence to improve pollution regulation in India3
Cause: {'name': 'Independent audits with predetermined payments', 'sent_id': 2, 'pos': [0, 5], 'type': 'CAUSE'}
Effect: {'name': 'on-site rechecking of audit data on a random basis', 'sent_id': 2, 'pos': [6, 17], 'type': 'CAUSE'}


Title: Using evidence to improve pollution regulation in India3
Cause: {'name': 'on-site rechecking of audit data on a random basis', 'sent_id': 2, 'pos': [6, 17], 'type': 'CAUSE'}
Effect: {'name': 'paying their own auditors', 'sent_id': 1, 'pos': [13, 17], 'type': 'CAUSE'}


Title: Using evidence to improve pollution regulation in India3
Cause: {'name': 'on-site rechecking of audit data on a random basis', 'sent_id': 2, 'pos': [6, 17], 'type': 'CAUSE'}
Effect: {'name': 'corruption', 'sent_id': 1, 'pos': [18, 19], 'type': 'EFFECT'}


Title: Using evidence to improve pollution regulation in India3
Cause: {'name': 'on-site rechecking of audit data on a random basis', 'sent_id': 2, 'pos': [6, 17], 'type': 'CAUSE'}
Effect: {'name': 'misreporting of industrial emissions', 'sent_id': 1, 'pos': [20, 24], 'type': 'EFFECT'}


Title: Using evidence to improve pollution regulation in India3
Cause: {'name': 'more accurate information', 'sent_id': 2, 'pos': [18, 21], 'type': 'EFFECT'}
Effect: {'name': 'corruption', 'sent_id': 1, 'pos': [18, 19], 'type': 'EFFECT'}
```

Figure 4.4: Predictions(ATLOP/CREST1)



```
Title: Using evidence to improve pollution regulation in India3
Cause: {'name': 'existing arrangement of industrial firms’ hiring', 'sent_id': 1, 'pos': [5, 12], 'type': 'CAUSE'}
Effect: {'name': 'corruption', 'sent_id': 1, 'pos': [18, 19], 'type': 'EFFECT'}


Title: Using evidence to improve pollution regulation in India3
Cause: {'name': 'existing arrangement of industrial firms’ hiring', 'sent_id': 1, 'pos': [5, 12], 'type': 'CAUSE'}
Effect: {'name': 'misreporting of industrial emissions', 'sent_id': 1, 'pos': [20, 24], 'type': 'EFFECT'}


Title: Using evidence to improve pollution regulation in India3
Cause: {'name': 'Independent audits with predetermined payments', 'sent_id': 2, 'pos': [0, 5], 'type': 'CAUSE'}
Effect: {'name': 'more accurate information', 'sent_id': 2, 'pos': [18, 21], 'type': 'EFFECT'}


Title: Using evidence to improve pollution regulation in India3
Cause: {'name': 'on-site rechecking of audit data on a random basis', 'sent_id': 2, 'pos': [6, 17], 'type': 'CAUSE'}
Effect: {'name': 'more accurate information', 'sent_id': 2, 'pos': [18, 21], 'type': 'EFFECT'}


Title: Using evidence to improve pollution regulation in India3
Cause: {'name': 'Independent audits with predetermined payments', 'sent_id': 2, 'pos': [0, 5], 'type': 'CAUSE'}
Effect: {'name': 'prompted firms to lower pollution', 'sent_id': 2, 'pos': [22, 27], 'type': 'EFFECT'}


Title: Using evidence to improve pollution regulation in India3
Cause: {'name': 'on-site rechecking of audit data on a random basis', 'sent_id': 2, 'pos': [6, 17], 'type': 'CAUSE'}
Effect: {'name': 'prompted firms to lower pollution', 'sent_id': 2, 'pos': [22, 27], 'type': 'EFFECT'}


Title: Using evidence to improve pollution regulation in India3
Cause: {'name': 'Independent audits with predetermined payments', 'sent_id': 2, 'pos': [0, 5], 'type': 'CAUSE'}
Effect: {'name': 'false reports of compliance with emissions norms reduced by 80 per cent', 'sent_id': 3, 'pos': [9, 21], 'type': 'EFFECT'}


Title: Using evidence to improve pollution regulation in India3
Cause: {'name': 'on-site rechecking of audit data on a random basis', 'sent_id': 2, 'pos': [6, 17], 'type': 'CAUSE'}
Effect: {'name': 'false reports of compliance with emissions norms reduced by 80 per cent', 'sent_id': 3, 'pos': [9, 21], 'type': 'EFFECT'}


Title: Using evidence to improve pollution regulation in India3
Cause: {'name': 'existing arrangement of industrial firms’ hiring', 'sent_id': 1, 'pos': [5, 12], 'type': 'CAUSE'}
Effect: {'name': 'false reports of compliance with emissions norms reduced by 80 per cent', 'sent_id': 3, 'pos': [9, 21], 'type': 'EFFECT'}


Title: Using evidence to improve pollution regulation in India3
Cause: {'name': 'paying their own auditors', 'sent_id': 1, 'pos': [13, 17], 'type': 'CAUSE'}
Effect: {'name': 'corruption', 'sent_id': 1, 'pos': [18, 19], 'type': 'EFFECT'}
```

Figure 4.5: Predictions(ATLOP/IE)

4.5 Discussion

In this section, we discuss our results for the experiment.

Starting with the Bi-LSTM-CRF model, the low scores on both CREST variations are indicators for the model not working for these datasets. As we have seen from the example predictions, the model predictions are limited in length. The prediction of 'effectiveness', while can be seen as a causal candidate on its own in some cases, does not include the full meaning of the original cause span. One major reason for this appears to be the difference in the labeling schemes of the CREST and IE datasets. Due to shorter and syntactically more simple examples available in the CREST collections, the predictions are performed poorly and the scores obtained are very low.

The observations from the loss graphs for CREST-1 and CREST-2, where the validation loss starts stagnating quickly and then increases, while the training loss is decreasing, are signs that the model is not fitting for the test set. This also supports the above mentioned problem of difference in datasets.

Another aspect worth mentioning is that the additional samples containing the medical datasets [16], [32] and the SemEval 2020 Task 5 [57] have not increased the scores noticeably: only the precision score has increased by 3% for the effect labels. We can conclude that these mentioned datasets do not include samples that allow an improvement for the causality extraction in the impact evaluations domain.

Another aspect that draws an eye is the fact that the prediction of effect candidates has higher precision scores than the cause candidates in all datasets. Two possible reasons come to mind on this occurrence. It was mentioned before that the intervention names in several cases consist of unusual, foreign words, and also in some cases abbreviations. As the CREST variations are not containing any kind of examples for any specific intervention names, as it is a general dataset for causality, it can cause the scores for the cause labels to be lower than effect labels. In case of the IE dataset, new words for intervention names and a lack of data might be reasons for the low scores on the Cause label. The second reason would be the effect spans that are longer and that contain verbs with causal meaning. As seen from our labeling scheme in our methodology, it is very likely that the model can predict the verbs inside the effect spans, which causes higher scores on effect labels in general. It is possible because the longer effect spans contain more words in general, more predictions are made on tokens that correspond to an effect span.

The scores for the IE dataset, while standing relatively higher than the CREST variations, are considerably low. When the IE dataset is used, longer spans are more likely to be predicted, as we have seen from the example predictions and the scores. The prediction can be considered as correctly done, as it mostly conserves the meaning of the full span. However, as seen in the second example, it also could not predict the intervention. Following these, one cause for the low scores appears to be the low amount of data used for the training. Considering the amount of papers for training we have used, the amount of interventions labeled is also low. This would be a reason for the false prediction of the second example.

We conclude that the use of the Bi-LSTM-CRF with this implementation is not sufficient for both of the cases.

Looking at the RE models, we can see that the models are relatively inefficient when they are trained with CREST datasets compared to when they are trained with our impact evaluations dataset. The relatively unstable recall scores and very low scores on the other metrics indicate that these general datasets are not suitable for the relation extraction task in our model. The relation predictions made by use of CREST also show the unstability of the predictions, where a cause is linked to another cause entity candidate. These occurrences can be attributed to the following: The variations of CREST are not enough for the relation extraction, and the labeling scheme difference causes the low scores.

The losses for DocuNet, when considering only the given cross-validation sets, also imply overfitting like the case in NER, which again supports the problem being the difference in the datasets. It is also possible that the directly increasing validation loss might be caused by the validation set not being shuffled randomly enough, validation and training losses of the other cross-validation sets should be inspected to reach a full conclusion.

On the other hand, we can infer from the results of the evaluation with the IE dataset that the relation extraction models are performing well considering the size of the dataset. With only 237 sample documents and less than 1000 example relations, the improvement in the scores from CREST to IE dataset is considerably successful.

5. Conclusion and Future Work

This chapter summarizes our work and covers several future work possibilities.

5.1 Conclusion

To sum up our work, we have tried to design a framework that aimed to extract causality in the specific field of impact evaluations. For this, we have combined the tasks of Named Entity Recognition and Relation Extraction into our framework, using several state-of-the-art models available. We can conclude that as a whole, our framework is not able to extract the interventions and their effects, especially when using a general domain dataset. In its current state our scheme is not a reliable method for the extraction of interventions and their effects. Our first technique to extract entity candidates of cause and effects performs poorly on general data. On the domain specific IE dataset, it also performs relatively poorly. The second approach to extract relations between found candidates performs unstable and poorly on the general data, but gives promising results on the domain-specific dataset despite low amounts of data. The poor performance of the NER section of the model limits the end results of the whole framework. While the interventions and their effects can be extracted for some cases with the domain-specific IE dataset, the scores and predictions are worse than desired.

The main reasons for the failure of the model appear to be the differences in the general data and the domain specific data, and a general lack of data in the domain of impact evaluations.

5.2 Future Work

There are five aspects that come to mind regarding the possible future works:

The most important point that must be addressed in the future is definitely the lack of data, not only in the domain of impact evaluations, but also in many other domains. The intervention corpus of Min et al.[33] can be incorporated for extraction of interventions in the future.

The second aspect to focus on is the extraction of causal candidates. The extraction of causal candidates can be done in a different approach. The implementation of Bi-LSTM-CRF can be improved by providing Part-of-Speech tags, similar to DeepCause [36]. The intervention extraction can also be separated from the causal candidate extraction of our work. An example for this can be through incorporating a similar method like in Min et al.'s intervention extraction scheme [33].

Another major aspect for the future works to focus on should be an unanimous annotation scheme for causal relations. The longer sequences of spans used in our labeling scheme and the shorter spans available in the general CREST dataset show that the data can be annotated differently.

One other aspect is the other causality types that we have not considered in our work, like the chained causality we have mentioned. There are works that also consider this type of causality [7], and it could be incorporated in the field of impact evaluations in the future.

Final things to consider are the remaining problems in our work. These are adding the remaining results from the ATLOP and DocuNet loss values, and the missing training losses on the DocuNet reports for IE dataset. The tokenization also still remains a problem for the CREST-1 variation we have created, which can be looked into in a future work. A scheme to distingush general causal entities in the IE dataset after the change of ”Intervention” labels into ”Cause” labels can also be considered as a future work.

Bibliography

[1] Bengt Altenberg. “Causal linking in spoken and written English”. In: *Studia linguistica* 38.1 (1984), pp. 20–69.

[2] Nabiha Asghar. “Automatic extraction of causal relations from natural language texts: a comprehensive survey”. In: *arXiv preprint arXiv:1605.07895* (2016).

[3] Manvi Breja and Sanjay Kumar Jain. “Causality for Question Answering.” In: *COLINS*. 2020, pp. 884–893.

[4] Quoc-Chinh Bui et al. “Extracting causal relations on HIV drug resistance from literature”. In: *BMC bioinformatics* 11 (2010), pp. 1–11.

[5] Tommaso Caselli and Piek Vossen. “The Event StoryLine Corpus: A New Benchmark for Causal and Temporal Relation Extraction”. In: *Proceedings of the Events and Stories in the News Workshop*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 77–86.

[6] *Causal Knowledge Extraction and Assembly Toolkit*. https://worldmodelers.com/reading-assembly.html. Accessed: 25 March 2023.

[7] Tirthankar Dasgupta et al. “Automatic Extraction of Causal Relations from Text using Linguistically Informed Deep Neural Networks”. In: *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 306–316.

[8] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[9] Quang Do, Yee Seng Chan, and Dan Roth. “Minimally Supervised Event Causality Identification”. In: *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Edinburgh, Scotland, July 2011.

[10] Jesse Dunietz, Lori Levin, and Jaime Carbonell. “The BECauSE Corpus 2.0: Annotating Causality and Overlapping Relations”. In: *Proceedings of the 11th Linguistic Annotation Workshop*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 95–104.

[11] *Evidence impact summaries*. https://www.3ieimpact.org/evidence-hub/evidence-impact-summaries. Accessed: 23 March 2023.

[12] Paul J Gertler et al. *Impact evaluation in practice*. World Bank Publications, 2016.

[13] John Giorgi, Gary Bader, and Bo Wang. “A sequence-to-sequence approach for document-level relation extraction”. In: *Proceedings of the 21st Workshop on Biomedical Language Processing*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 10–25.

[14] Roxana Girju et al. “SemEval-2007 Task 04: Classification of Semantic Relations between Nominals”. In: *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 13–18.

[15] Alex Graves and Jürgen Schmidhuber. “Framewise phoneme classification with bidirectional LSTM and other neural network architectures”. In: *Neural networks* 18.5-6 (2005), pp. 602–610.

[16] Harsha Gurulingappa et al. “Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports”. In: *Journal of biomedical informatics* 45.5 (2012), pp. 885–892.

[17] Zhiyong He et al. “A survey on recent advances in sequence labeling from deep learning models”. In: *arXiv preprint arXiv:2011.06727* (2020).

[18] Stefan Heindorf et al. “CauseNet: Towards a Causality Graph Extracted from the Web”. In: *CIKM*. ACM, 2020.

[19] Iris Hendrickx et al. “SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals”. In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 33–38.

[20] Matthew Honnibal and Ines Montani. “spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing”. To appear. 2017.

[21] Pedram Hosseini, David A Broniatowski, and Mona Diab. “Predicting Directionality in Causal Relations in Text”. In: *arXiv preprint arXiv:2103.13606* (2021).

[22] Zhiheng Huang, Wei Xu, and Kai Yu. “Bidirectional LSTM-CRF models for sequence tagging”. In: *arXiv preprint arXiv:1508.01991* (2015).

[23] International Initiative for Impact Evaluation (3ie). *Impact Evaluation*. https://www.3ieimpact.org/What-we-offer/impact-evaluation. Accessed: 15 March 2023.

[24] International Initiative for Impact Evaluation (3ie). *International Initiative for Impact Evaluation*. https://www.3ieimpact.org. Accessed: 15 March 2023.

[25] *Improving targeting in social welfare programmes in Indonesia*. https://www.3ieimpact.org/evidence-hub/Evidence-impact-summaries/improving-targeting-social-welfare-programmes-indonesia. Accessed: 23 March 2023.

[26] jidasheng. *bi-lstm-crf*. https://github.com/jidasheng/bi-lstm-crf. Accessed: 18 March 2023.

[27] CHRISTOPHER S. G. KHOO et al. “Automatic Extraction of Cause-Effect Information from Newspaper Text Without Knowledge-based Inferencing”. In: *Literary and Linguistic Computing* 13.4 (Dec. 1998), pp. 177–186. ISSN: 0268-1145. eprint: https://academic.oup.com/dsh/article-pdf/13/4/177/10888761/177.pdf.

[28] Christopher SG Khoo, Syin Chan, and Yun Niu. "Extracting causal knowledge from a medical database using graphical patterns". In: *Proceedings of the 38th annual meeting of the association for computational linguistics*. 2000, pp. 336–343.

[29] John Lafferty, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". In: (2001).

[30] Jing Li et al. "A survey on deep learning for named entity recognition". In: *IEEE Transactions on Knowledge and Data Engineering* 34.1 (2020), pp. 50–70.

[31] Shining Liang et al. "A multi-level neural network for implicit causality detection in web texts". In: *Neurocomputing* 481 (2022), pp. 121–132.

[32] Claudiu Mihăilă et al. "BioCause: Annotating and analysing causality in the biomedical domain". In: *BMC bioinformatics* 14 (2013), pp. 1–18.

[33] Bonan Min et al. "Towards Machine Reading for Interventions from Humanitarian-Assistance Program Literature". 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6444–6448.

[34] Paramita Mirza et al. "Annotating Causality in the TempEval-3 Corpus". In: *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*. Gothenburg, Sweden: Association for Computational Linguistics, Apr. 2014, pp. 10–19.

[35] Nasrin Mostafazadeh et al. "CaTeRS: Causal and Temporal Relation Scheme for Semantic Annotation of Event Structures". In: *Proceedings of the Fourth Workshop on Events*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 51–61.

[36] Roland Mueller and Sardor Abdullaev. "DeepCause: Hypothesis Extraction from Information Systems Papers with Deep Learning for Theory Ontology Learning". In: Jan. 2019.

[37] Roland Mueller and Sebastian Huettemann. "Extracting Causal Claims from Information Systems Papers with Natural Language Processing for Theory Ontology Learning". In: Jan. 2018.

[38] Rutu Mulkar-Mehta et al. "Causal markers across domains and genres of discourse". In: *Proceedings of the sixth international conference on Knowledge capture*. 2011, pp. 183–184.

[39] David Nadeau and Satoshi Sekine. "A survey of named entity recognition and classification". In: *Lingvisticae Investigationes* 30.1 (2007), pp. 3–26.

[40] Hiroki Nakayama et al. *doccano: Text Annotation Tool for Human*. Software available from https://github.com/doccano/doccano. 2018.

[41] Qiang Ning et al. "Joint Reasoning for Temporal and Causal Relations". In: *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 2278–2288.

[42] phosseini. *CREST: A Causal Relation Schema for Text.* https://github.com/phosseini/CREST. Accessed: 15 March 2023.

[43] Rashmi Prasad et al. *Penn Discourse Treebank Version 3.0.* Version V1. 2019.

[44] Oxford University Press. *Definition of causality.* https://www.lexico.com/definition/causality. Accessed: 23 March 2023.

[45] PyTorch. *Advanced: Making Dynamic Decisions And The Bi-LSTM CRF.* https://pytorch.org/tutorials/beginner/nlp/advanced_tutorial.html. Accessed: 15 March 2023.

[46] Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. "Learning Causality for News Events Prediction". In: *Proceedings of the 21st International Conference on World Wide Web.* WWW '12. Lyon, France: Association for Computing Machinery, 2012, 909–918. ISBN: 9781450312295.

[47] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. "Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning." In: *AAAI spring symposium: logical formalizations of commonsense reasoning.* 2011, pp. 90–95.

[48] Yutaka Sasaki. "The truth of the F-measure". In: *Teach Tutor Mater* (Jan. 2007).

[49] Alisa Smirnova and Philippe Cudré-Mauroux. "Relation extraction using distant supervision: A survey". In: *ACM Computing Surveys (CSUR)* 51.5 (2018), pp. 1–35.

[50] Yifan Sun et al. "Circle loss: A unified perspective of pair similarity optimization". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 2020, pp. 6398–6407.

[51] Judith Jarvis Thomson. "Verbs of action". In: *Synthese* (1987), pp. 103–122.

[52] *World Modelers.* https://www.darpa.mil/program/world-modelers. Accessed: 25 March 2023.

[53] Ye Wu et al. "Renet: A deep learning approach for extracting gene-disease associations from literature". In: *Research in Computational Molecular Biology: 23rd Annual International Conference, RECOMB 2019, Washington, DC, USA, May 5-8, 2019, Proceedings 23.* Springer. 2019, pp. 272–284.

[54] Benfeng Xu et al. "Entity Structure Within and Throughout: Modeling Mention Dependencies for Document-Level Relation Extraction". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.16 (2021), pp. 14149–14157.

[55] Jinghang Xu et al. "A Review of Dataset and Labeling Methods for Causality Extraction". In: *Proceedings of the 28th International Conference on Computational Linguistics.* Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 1519–1531.

[56] Jie Yang, Soyeon Caren Han, and Josiah Poon. *A Survey on Extraction of Causal Relations from Natural Language Text.* 2021. arXiv: 2101.06426 `[cs.IR]`.

[57] Xiaoyu Yang et al. "SemEval-2020 Task 5: Counterfactual Recognition". In: *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020).* Barcelona, Spain, 2020.

[58] Liuyi Yao et al. “A Survey on Causal Inference”. In: *ACM Trans. Knowl. Discov. Data* 15.5 (2021). ISSN: 1556-4681.

[59] Yuan Yao et al. “DocRED: A Large-Scale Document-Level Relation Extraction Dataset”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 764–777.

[60] Shuang Zeng et al. “Double Graph Based Reasoning for Document-level Relation Extraction”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020, pp. 1630–1640.

[61] Ningyu Zhang et al. “Document-level Relation Extraction as Semantic Segmentation”. In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. Ed. by Zhi-Hua Zhou. Main Track. International Joint Conferences on Artificial Intelligence Organization, Aug. 2021, pp. 3999–4006.

[62] Wenxuan Zhou et al. “Document-level relation extraction with adaptive thresholding and localized context pooling”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 16. 2021, pp. 14612–14620.

A. Appendix

A.1 Discarded Methods

This section explains another methods we have tried to incorporate for the causality extraction and failed to do so.

The framework MCDN(Multi-level Causality Extraction Network)[31] was considered to be used in the beginning of the research phase. Due to the model being structured for sentence-level causality extraction, the model was discarded. The model uses multi-head self-attention as a baseline. Main focus of the work is to infer implicit causality using web articles, where they concatenate the word and segment level representations of sentences and extract causality [31].