UNIVERSITÉ LIBRE DE BRUXELLES, UNIVERSITÉ D'EUROPE

ULB

Solvay Brussels School
Economics & Management

MASTER THESIS

Presented with a view to obtain the Master in
Economics, finality Economic Analysis and European Policy

Automated extraction of causal diagrams from scientific literature in economics

By Valérie Devos

Director: Professor Hugues Bersini
Assessor: Professor Dimitris Sacharidis

Academic year 2021 - 2022

I authorize the consultation of this thesis

Acknowledgments

I would like to thank all the people who have supported me through the writing of this thesis. And all my friends for supporting me through the years and making me grow. I would like to thank Mr. Hugues Bersini, my thesis director who advised me throughout this work. A special thank to the people that corrected this work namely Alberto Zorzato and Dominique D'heedene. Moreover, I would like to express my gratitude to my family and especially to my parents who have encouraged me during the last 5 years.

Executive summary

Causal diagrams and their use in causal inference were developed mainly by **Judea Pearl**. He argues that they are a **useful mathematical language** to think about causality and a great tool for knowledge representation. The creation of a causal diagram using natural language processing on economic literature is examined in this thesis.

Section 2 goes through some theory assessing how we can infer causality from causal diagrams. Firstly, **graph terminology** and characteristics are discussed. Secondly, this thesis examines how to infer **conditional independency** between variables using the topology of diagrams. Conditional independency is used to confirm the exogeneity assumption in regression models. Moreover, the **d-separation** criterion is introduced, which is used to block backdoor paths that create spurious correlations between the independent and dependent variables. Then, we examine how we can use causal diagrams to tackle the **confounding bias**, namely by using random controlled trials (RCT's), the **backdoor and frontdoor criteria**. Furthermore, the **do-operator**, which give a causal language to statistics and gives the possibility to differentiate between observational data and experimental data is introduced. The **do-calculus** is discussed which gives a number of steps to follow to transform an equation containing a do-operator, by a do-free equation. The do-calculus is a generalization of the backdoor and frontdoor criteria and is used to find out if we can infer a causal relationship between 2 variables with observational data or if we need an experiment. It uses a causal diagram as input.
The main arguments for the use of causal diagrams are: 1) They are a **clear and transparent representation** of how the world works. 2) They use a **causal language** contrary to statistics. 3) If we assume the diagram to be a representation of how the world works, **causal inference can be done based on them.** 4) They are useful to uncover which **confounding variable** we should consider. 5) They are useful in settings where a lot of variables interact. However, in recent years, their use has been fiercely **debated in economics**. Indeed, some economists argue that their use could **mislead researchers** and that diagrams do not bring much more than what econometrics already does. **Guido Imbens** for example claims that causal diagrams are not suited to integrate some assumptions like monotonicity or convexity, that the potential outcome framework has long established the inference strategies needed and that **causal diagrams lack of convincing empirical applications**. On the other hand, other economists wonder why they have been **used extensively in other fields** like epidemiology and not in economics.

In section 3, the 3 ways of constructing a causal diagram are examined. Firstly, it can be **done manually** by the researcher. After reading several papers on a certain topic the researcher draws the relationships between variables. Secondly, one can use **numeric data and an algorithm** (PC algorithm). By using graph theory and **conditional independencies** the algorithm can create a causal diagram from numeric data. Sometimes this method results in a partially directed graph that should be completed by intuition. The third way is **automatically extracting information from literature**. This is the newest method and it uses **Natural Language Processing** (NLP) to extract information from unstructured text and represents it in the form of a causal diagram. This last method is the one that is going to be developed in this thesis.

Section 4 reviews the existing **literature of automated information extraction** in economics and other fields. The development of **natural language processing** in recent years is discussed. At first, information extraction relied on **rule-based** models which included a complex set of rules to find patterns. Later on, **statistical-based** methods emerged using **machine learning models** which can be trained by a set of labelled documents. **Causality extraction** has a specific literature. Indeed, causal relationships can appear in many syntactic forms and are classified into different categories (explicit, implicit, inter and intra-sentential). Moreover, causality can be represented by many different kinds of connectives (verbs, adjective, adverbs, etc). Because of the many different forms causal language can take, machine learning is the best method to extract causal relationships. Indeed rule-based models would be too labor-intensive and would not extract implicit causal links. The application of automatic causality extraction on scientific literature has mainly been done in the life-sciences literature (medical sector, biology epidemiology, etc.). In medicine, for example, causal relationships are important because they are used to determine if a medicine will cause the condition of a patient to improve. Automatic causality extraction is **underrepresented in economic literature**. Most of the articles concerning this topic in economics were written recently. The articles are discussed and compared to the model that is created in section 6.

In section 5, the motivations and potential applications of a model that automatically extracts causality in economics are discussed. 1) The **growing number of papers** makes reading increasingly time consuming for researchers. When researchers have to write an article, they have to read through a lot of articles, maybe some information gets lost, this could be fixed by an automated method. 2) The **recent debate** about their use makes their study valuable. Indeed, it is easier to prove the usefulness of a tool when the access to the tools is simple 3) Causal discovery algorithms often produce partially directed graph. Graphs made by numeric data and graphs made by text data could be combined to solve this issue and yield a completely directed graph. 4) The method is used successfully in **other fields.** 5) It is a **tool for variable selection** (tackling the confounding bias) and detection 6) It makes **assumptions more transparent** and auditable which could increase **the spread** of the knowledge. One can use a selection diagram to address external validity. 7) Create **an artificial researcher**. A **global causal diagram** could be used to automatically generate questions and hypothesis.

In section 6, an automated causal extraction model is created. IBM Watson is one of the most advanced tools for Natural Language Processing (NLP) and offers a simple interface that does not need too much coding. This also means that it is accessible for economists. IBM provides free access to some of their tools under the "*IBM academic initiative program"*. Therefore, **IBM Watson tools are used to train a supervised learning model**, to recognize causal relations from scientific abstracts in a specific economic domain. Supervised learning is the most common method to analyze domain specific language, this is why it is suited for economic literature. The goal is to extract **findings** from the papers. 2 services of IBM Watson are used, namely **Watson Knowledge Studio** (WKS) to annotate abstract and create the model and **Watson Natural Language Understanding** (NLU) to deploy the model created in WKS and construct a causal diagram. **102 abstracts discussing "income inequality" were annotated** to

train the model. The annotations contain the name of **economic variables**, relations words and evidence words as entity types. 3 different kind of relations between variables are considered: **"causality", "correlation" and "no relation"**. The documents were pre-annotated by using a dictionary containing economic variables. After training the model on a small set of documents it was used to pre-annotate the other documents. This reduced considerably the time needed to label all the abstracts. The main annotation difficulties were: the size of economic variables that contain often more than one word, the annotation of conditional causality and the ambiguity of causal language. Annotation guidelines were created to have a homogeneous training set. 70% percent of the documents was used to train the model 23% was allocated to the test set and 7% was allocated to the blind set. Different versions of the model were created, they differ by number of annotated documents, the number of entries in the dictionary, the type system and the consistency of annotations. The best version is version 9 which yielded an **F1 score of 0.83 for entity recognition and 0.42 for relation recognition** in the test set. The blind set yielded comparable results. This is an increase of 0.32 for entity recognition and 0.15 for relation recognition in comparison with version 1. The main failures of the model reside in: failing to find relationships because it fails to identify correctly the entities in the first place, mixing causal relation and correlation and relations that are annotated by the model and that do not represent findings of the paper. The low F1 score for relations extraction is not surprising considering the **small amount of training data** with in total around 13.000 words in the training set while the WKS documentation recommends having 300.000 words in the training set to yields sufficient results.

In section 7, version 9 of the model is deployed to the NLU service and used to create a causal diagram. 23 abstracts present in the test set were used to extract causal relationships. The model finds 76 relations (cause – relation – effect). After deleting doubles, 66 unique relations are left. All the relations come with a probability score. Only the relations that have a probability higher than 50% to be correct are represented in the causal diagram, this is 56 out of the 66 triplets. The nodes that are the most connected are "Income inequality" and "Inequality", this is because most of the articles discuss this topic.

This thesis shows that it is possible to extract causal relationships from abstracts in economics with a supervised machine learning model. The accuracy is low for the relation extraction, but we can assume that with a bigger training set the results would be more accurate. Some hypotheses are made to answer why causal diagrams are still underexploited in economics. Finally, ideas for future research are given.

Table of content

Table of figures

Table of tables

1. Introduction

The use of causal diagrams for inferring causal relationships has been mostly developed by **Judea Pearl**. He claims that such diagrams give an intuitive representation of how the world works and that they rely on causal language, which is not the case in statistics (Pearl, 1995; Pearl, 2016; Pearl, 2018). However, causal diagrams are absent in economic textbooks, which prefer to focus on statistical tools like the potential outcome framework (Imbens, 2020). Pearl urges economists to use causal diagrams, which have been found to be useful in other domains (Pearl, 2014a). Imbens, on the other hand, is not convinced of the method and argues that it lacks convincing empirical applications (Imbens, 2020). The debate about the use of diagrams for causal inference in economics does often not take into account the different ways of constructing them. If we were to be able to construct those diagrams in a reliable and fast way, this would maybe change the outcome of the debate.

In recent years, we have seen an exponential increase of scientific literature in all fields (White, 2017). The main way of exploiting this unstructured knowledge for researchers is to manually review this literature. This process is time-consuming and labor intensive. Indeed, researchers spend most of their time on knowledge extraction and yet, they are not able to read all the literature out there. With the rise of natural language processing and machine learning, we now have the opportunity to extract information automatically from articles. Ontologies in scientific literature mostly link authors to their papers, publisher, years, organizations, research topics, etc. They include almost no information about the content of the articles (Chen, 2019; Dessi, 2021). Including findings in those ontologies is difficult because it requires to extract causal relationships from articles which can be expressed in many different ways. Automatic causality extraction has been done extensively in life sciences but has not been fully exploited in economics (Chen, 2020).

This **thesis aims to automatically extract findings in the form of causal relationships from abstracts in economic literature**. Then, to represent the extracted information in the form of causal diagrams, which provide a useful representation to think about causality.

In section 2, the **theory of causal diagrams** and their use in the area of causal discovery is discussed. In section 3, the **different ways of creating causal diagrams** are examined. In section 4, the existing **literature of automated information extraction** in economics and other fields is discussed. In section 5, the **motivation and potential applications** for automated causal diagrams from scientific literature are highlighted. In section 6, **IBM Watson** tools are used to train a **machine learning model** on economic literature. Different versions of the model are discussed, and the accuracy of the model is examined. In section 7, the model created in section 6 is used to create a causal diagram. Finally, the potential **issues that are preventing causal diagrams to be fully exploited** in economics are discussed and several ideas for further work are given.

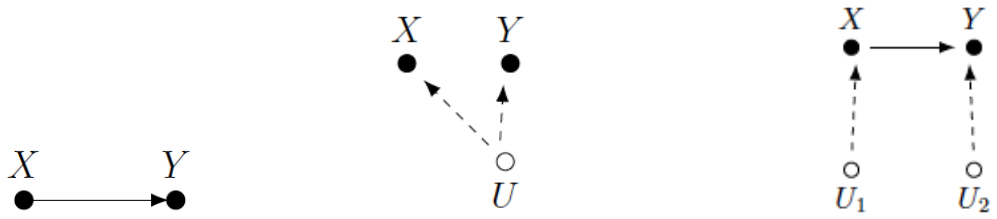
2. Causal inference with causal diagrams: Theory and discussion

In this section some **graph terminology** will be introduced and its use for causal inference will be examined. We will assess how causal diagrams can be used to infer conditional independence between variables and tackle the **confounding bias**. The **do-calculus** will be introduced. Finally, we will discuss the **recent debate about their use** in economics.

> *"Graph theory provides a **useful mathematical language** to think about causality, it is seen as an attractive way to capture how people think about causal relationships. It allows to check the validity of causal statements based on **intuitive graphical criteria**, that do not require any algebra. It represents an underlying structural causal model. In addition, they open up the possibility to completely **automate the causal inference** task with the help of special identification algorithms"* (Hünermund, 2021).

2.1. Some graph terminology

Causal diagrams are referred to in different ways depending on the literature of interest. In this thesis we will refer to causal diagrams as graphs used for causal discovery, which were put into light mainly by **Judea Pearl,** who dedicated more than 30 years of research on the topic. In the literature, we find several denominations: causal diagrams, knowledge graphs, **directed acyclic graphs** (DAG), directed graphical causal models (DGCM), etc. They all represent the same thing. In short, we speak about graphs including **nodes which represent variables** and **directed edges** between those nodes representing the relation between those variables. If we consider an experiment that looks for the causal link of variable X on variable Y and we assume that no other variable has an impact on neither X nor Y, then the causal diagram will be denoted by Figure 1a.

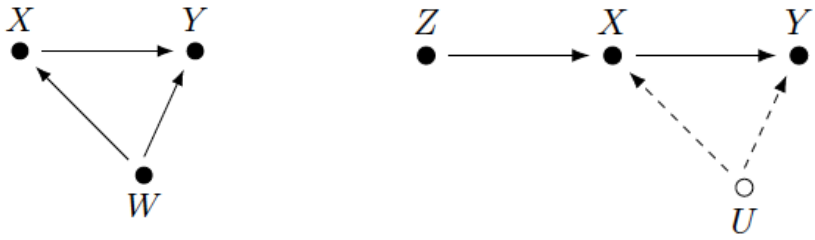


a) X caused Y b) Spurious correlation c) Alternative graph

Figure 1: causal diagrams with 2 observed variables based on (Imbens, 2020)

Because causal relations are asymmetric, directed edges will be used. Indeed, if X has a causal relation on Y, then it is not possible for Y to have a causal relation on X. The diagram captures causal relations, but it also captures the absence of causal relations if there is no edge between variables (Hünermund, 2021).

Figure 1b, an unobserved variable U is causing both X and Y, causing a **spurious correlation** between X and Y. An observed variable is denoted by a **plain node** while an unobserved variable is denoted by an **empty node** U. The relation between an unobserved variable and another variable (observed or unobserved) is represented by a **dashed edge** while 2 measured quantities are measured by **plain edges**. Figure 1c is an alternative way of representing Figure 1a: given that variable U1 and variable U2 have no relation between each other, their presence does not influence the interpretation of the causal relation between X and Y, so we can omit them from the graph by convention. In Figure 1, we have seen a simple model with 2 variables, but more complicated setups are possible.



a) Confounder b) Instrumental variables

Figure 2: causal diagrams with 3 observed variables based on (Imbens, 2020)

In Figure 2a, 3 variables are observed, W is an observed **confounder** in this setting because it has an impact on X and Y. If we look only at the relation between X and Y without considering W we will end with a false causal effect between X and Y. This is called the **confounding bias**. However, we can avoid this bias and find the real causal relation between X and Y by adjusting or controlling for W, which means adding this variable in the regression. In Figure 2b, U is an unobserved confounder, so we cannot control for U. This makes it difficult to infer the causal relationship of X on Y. Fortunately the observed variable Z can help us tackle this problem, in this setting, Z is called an instrument. This is denoted as an **instrumental variable** setting (Angrist, 1996). In econometric' terminology X is said to be endogenous, as there is an unobserved confounder U affecting both X and Y. Z has no unobserved confounder impacting its relationship with the endogenous variable Y. Instrumental variables are used a lot in econometrics, but traditionally not represented in the form of causal diagrams (usually statistical form). Pearl (2018) argues that one useful thing about graphs is that they yield a **clear view of the assumptions** of the model, in the Figure 2b it is useful to represent the assumption of the instrumental variable setting.

One other advantage of DAGs is given by their structural meaning in comparison with econometrics. Indeed, in econometrics we use "=" to denote a causal relationship between X and Y, for example Y= aX, which is equivalent to X=Y/a. But, if X has a causal impact on Y, this does not necessarily mean that Y has a causal impact on X. Statistical language is not precise enough to speak about causality (Pearl, 2018). On the other hand, by using arrows "X←W" and "W←X" have a different meaning.

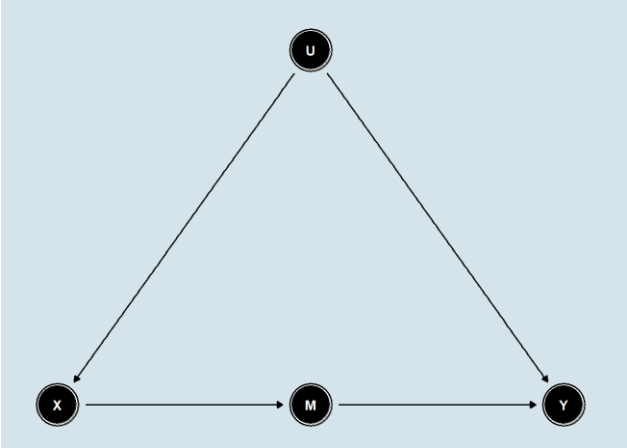


Figure 3 : Based on (Pearl, 2018)

In Figure 3, the **set of nodes** of the graph is S={X, M, Y, U} and the **edges** are X→M, M→Y, U→X and U→Y. The nodes which have arrows that are directed into other nodes are called **parents** of that node. Here X is a parent of M and M is a parent of Y. **Ancestors** of a node contain all the parents of that node and the parents of their parents. Here X, M and U are all the ancestors of Y. The nodes that have arrows directed into them from other nodes are called **children**. Here Y is a child of M. **Descendants** of a node contain all their children and the children of their children. Here M and Y are all the descendants of X. A **path** in a causal diagram is a succession of edges connecting 2 nodes disregarding the directions of the arrows. In Figure 3 we have 2 paths from X to Y: X→M→Y and X←U→Y. A path were all the arrows go in the same direction is called a **directed path**. A path from node X to node Y, beginning with an incoming arrow into X and ending with an incoming arrow into Y is called a **back-door path**. The path (X←U→Y) is a back-door path from X to Y (Pearl, 2000).

When we use the name directed acyclic graph, we use the word **acyclic** to point to graphs that contain **no feedback loop** (example of a feedback loop: X→M→Y→X), meaning that we do not consider variables that exert a causal influence upon themselves, which is similar to **recursive models** in econometrics. Indeed, only recursive models can give causal explanations (Maddala, 1986).

2.2. Causal diagrams and conditional independence

Conditional independence is an important concept in econometrics, it is related to the exogeneity assumption in regression models. In economics we care about the independence of the independent variable X and the dependent variable Y. Indeed, in an Ordinary Least Squared (OLS) regression, one of the assumptions stated by the Gauss-Markov theorem is strict exogeneity. Strict exogeneity means that the error term is uncorrelated with the independent variable X, $E[u_i | x_i] = 0$ (Stock, 2015). In other words, the independent variables X are not dependent on the dependent variable Y, meaning that the exogenous variable X has an impact on Y without being affected by Y. If this assumption is not true, X is considered as an **endogenous variable** and we have to adjust the model to find the true causal link.

> "D*irected acyclic graphs as economical schemes for representing conditional independence assumptions is well evidenced in the literature. Graphs are useful because they encode conditional independence relationships irrespective of the specific functional relationships between variables. We can thus* ***infer conditional independence relationships simply from the topology of the graph*** *(no need for tedious algebra).*" (Pearl, 1995).

Different types of paths are considered, and conditional independency is deduced between the variables in these paths.

> "*In a* ***chain****,* ***X→Z→Y****, X and Y are dependent, X is independent of Y conditional on Z (*$X \perp\!\!\!\perp Y|Z$*). In a* ***fork,*** ***X←Z→Y****, X and Y are dependent, but become independent conditional on Z (*$X \perp\!\!\!\perp Y|Z$*). In a* ***collider*** *setup* ***X→Z←Y****, if Z is a collider, X and Y are unconditionally independent, but become dependent conditional on Z (*$X \not\perp\!\!\!\perp Y|Z$*).*" (Hünermund, 2021).

Chain:	$A \to B \to C$	$\Rightarrow$	$A \not\perp\!\!\!\perp C$ and $A \perp\!\!\!\perp C \mid B$
Fork:	$A \leftarrow B \to C$	$\Rightarrow$	$A \not\perp\!\!\!\perp C$ and $A \perp\!\!\!\perp C \mid B$
Collider:	$A \to B \leftarrow C$	$\Rightarrow$	$A \perp\!\!\!\perp C$ and $A \not\perp\!\!\!\perp C \mid B$

Figure 4 : Recap of conditional independencies from (Hünermund, 2021)

The generalization of this theory to more than 3 variables is called the **d-separation criterion** (Geiger, 1990).

Definition: d-separation (Pearl, 2016)

"A path P is ***blocked*** *by a set of nodes B if and only if:*

1. *P contains a chain of nodes* ***X→Z→Y*** *or a fork* ***X←Z→Y*** *such that the middle node Z is in B (i.e., Z is conditioned on), or*
2. *P contains a collider* ***X→Z←Y*** *such that the collision node Z is not in B, and no descendant of Z is in B*"

If a path is not blocked it is said to be **open**. If we want to find the causal relationship between X and Y we have to block all the non-direct paths between them, the 2 nodes are then said to be d-separated which means they are independent. In other words, open paths create spurious correlations between X and Y. Those paths can be blocked by conditioning on certain intermediate variables on the path.

We will see in the next section how to handle confounding variables in more complicated settings.

2.3. Causal diagram to control for confounding bias

In Figure 2 we have seen how observed and unobserved confounders can create false causal relationships when not properly taken into consideration. One famous example of the confounding problem can be illustrated by the **Simpson's paradox** which is often found in social sciences and shows that statistics alone can lead to a false sense of causality. The paradox occurs when we find a causal relationship in a population and that this relationship disappears or reverse when we divide this population into several groups (Blyth, 1972). This can be fixed by taking confounding variables appropriately into account.

It is still difficult to find a clear definition of confounders, and to know if we should add them to the model or not. In 1996 a Norwegian epidemiologist Sven Hernberg said "*if you suspect a confounder, try to adjust for it and try not adjusting for it. If there is a difference, it is a confounder and you should trust the adjusted value. If there is no difference you are off the hook*" (Pearl, 2018). We speak about adjusting or controlling when we add the confounder into the regression. Pearl states that we shouldn't blindly adjust for all confounders and that this has misguided a century of economists, social scientists and epidemiologists. Moreover, confounding variables are known to be the main cause of the violations of the exogeneity assumption.

We have seen in different settings (chain, forks, colliders) how we can block paths with the d-separation, but what if we have a more complicated path like:

$$A \leftarrow B \leftarrow C \rightarrow D \leftarrow E \rightarrow F$$

Figure 5 : Complex backdoor path

Which variable should we control for in this case (Figure 5) if we want to find the causal impact of A on F? Controlling for all the variables would be a mistake, because the path contains a collider D, which implies that the backdoor path is already blocked. However, we can control for the other variables without opening the path, but this is not necessary.

Pearl argues that graphical methods have fully deconfounded the confounding problem. In his book he says:

> *"Although confounding has a long history in all areas of science, the recognition that the problem requires causal, not statistical solutions is very recent… There is now an almost universal consensus, at least among epidemiologists, philosophers and social scientists that 1. Confounding needs and has a causal solution 2. Causal diagrams provide a complete and systematic way of finding that solution.*" (Pearl, 2018).

We will see in the next section how causal diagrams can help us to **deconfound**. Through intervention, using the back-door or front-door criteria.

2.3.1. Causal diagrams of randomized controlled trials

In 1920, it became obvious that random experimental designs were the best way of quantifying a causal relationship, and randomized controlled trials (RCTs) became the gold standard. Causal diagrams are useful to understand why and how RCTs work. In Figure 6 we can see the effect of an RCT on a graph, it removes all the incoming arrows of the independent variable X. **Randomizing erases all the confounders** without injecting new ones. After intervention, there is no arrow going from Z to X and Z is not a confounder anymore (Figure 6). This is why RCTs are so popular and are useful to infer causal relationships. We could infer the causal relationship between X and Y by adjusting for Z, but one of the main advantages of RCTs is that they erase every confounding variable even those we couldn't think about or the ones we could not measure.

Figure 6 : Impact of an intervention on a graph (Hünermund, 2021)

In classical statistical language there is not a way to differentiate an experiment from an observation, P(Y|X=x) is used to represent both. To fix this, Pearl introduced the **do-operator** to represent interventions and give statistics a causal language. P(Y|do(X=x)) represent *"the probability distribution of Y if we fix (set) X to the specific value x."* while P(Y|X=x) represents the value of Y when we observe X=x (Pearl, 1995).

However, **intervention is not always feasible**. It is expensive and sometimes impossible or unethical. For example, we cannot randomly assign some kids to have a good education and others to have a bad education to observe the causal relation of education on wages. Or when inferring the impact of intelligence on wages, we cannot randomly assign people to be smart and others not to be. Researchers are often left with **observational data** to analyze. They have to be able to identify the effect of interventions from observational data only. They want to identify P(y|do(x)) but the only information they have is P(x,y,z). We will see how to transform equations containing the do-operator into do-free equations.

2.3.2. Backdoor criterion

In Figure 5, we have seen a complex backdoor path. It can happen that 2 variables (X, Y) are linked by several backdoor paths. The backdoor criterion relies on **blocking every backdoor path** by conditioning on some variables along the path to find the real causal effect between X and Y. The backdoor criterion identifies the minimum set of variables we have to condition on, to make sure that all backdoor paths between X and Y are blocked.

Definition: The **Backdoor Criterion** (Pearl, 2016)
"Given an ordered pair of variables (X, Y) in a directed acyclic graph G, a set of variables Z satisfies the backdoor criterion relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X. If a set of variables Z satisfies the backdoor criterion for X and Y, then the causal effect is given by:"

$$P(Y = y|do(X = x)) = \sum_{z} P(Y = y|X = x, Z = z)P(Z = z)$$

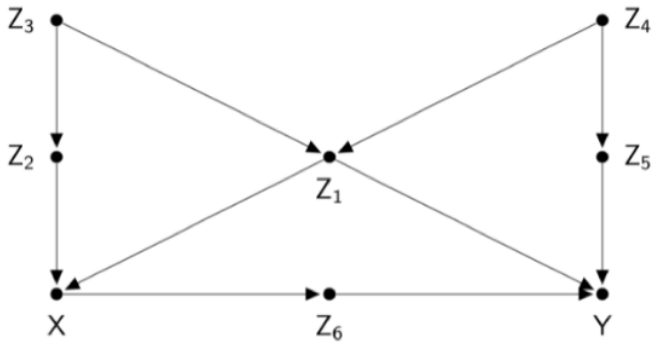


Figure 7 : Example for backdoor criterion (Hünermund, 2021)

Causal diagrams enable us to **easily know which variable we should adjust for** when facing the confounding bias, and to understand what the **different possible adjustment sets** are. To illustrate how the backdoor criterion works, we assume the causal relationships of Figure 7 to be correct. The goal is to determine the different **minimum adjustment sets** possible when looking for the causal relation between X and Y. First, we have to identify the different backdoor paths (see Figure 8).

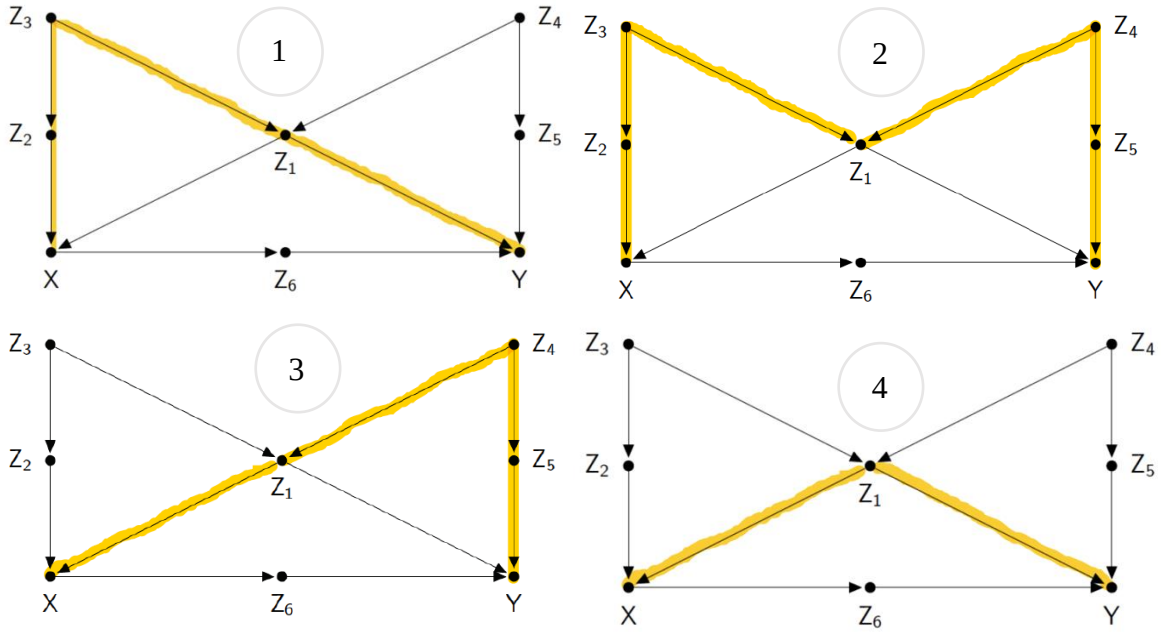


Figure 8: Different backdoor paths

Here we have **4 backdoor paths**, the shortest one is (path 4) X←Z1→Y, it is an open path because it is a fork, we can block it by adjusting for Z1. If we look at path 1 and 3, adjusting for Z1 also blocks those paths. This is not the case for path 2, X←Z2←Z3→Z1←Z4→Z5→Y. This path is a composition of a chain X←Z2←Z3, a fork Z2←Z3→Z1, a collider Z3→Z1←Z4, a fork Z1←Z4→Z5 and a chain Z4→Z5→Y. If we don't adjust for Z1 the whole path is considered as blocked because of the collider. If we adjust for Z1 the path is not blocked by the collider anymore and we open the path. But we have seen that we can block chains and forks by conditioning on their middle variable. So, we can choose to condition either on Z2, Z3, Z4 or Z5. This is why the minimum adjustment sets possible are {Z1; Z2}, {Z1; Z3}, {Z1; Z4}, {Z1; Z5}. Each of the different sets blocks all possible backdoor paths possible. In this example the adjustment set {Z1; Z2; Z3; Z4} also satisfies the backdoor criterion, but sometimes **it can be harmful to adjust for all the variables**. With the backdoor criterion we can find the different smallest adjustment set and choose the most convenient one, depending on the **available data**. Moreover, given that different adjustment sets are admissible, we should deduce the same causal effect with those different sets which can be useful to check the reliability of the data.

2.3.3. Frontdoor criterion

When we used the backdoor criterion, all the possible adjustment sets contained Z1. If we cannot observe Z1 we cannot apply the backdoor criterion because we cannot adjust for it. The **frontdoor criterion** is particularly useful when we are unable to observe some variables used as adjustment in the backdoor criterion case.

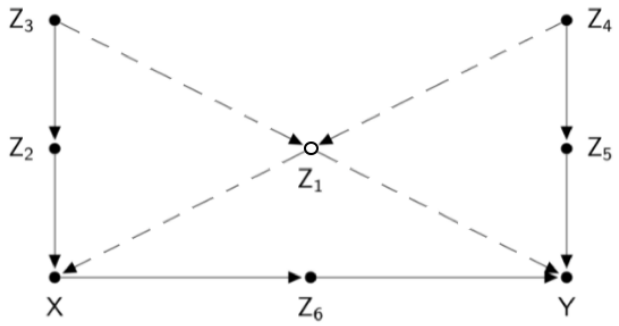


Figure 9: Example for frontdoor criterion (Hünermund, 2021)

Definition: The frontdoor criterion (Pearl, 2016)

"A set of variables Z is said to satisfy the frontdoor criterion relative to an ordered pair of variables (X, Y) if:

1. *Z intercepts all directed paths from X to Y*
2. *There is no unblocked path from X to Z*
3. *All backdoor paths from Z to Y are blocked by X*

If Z satisfies the frontdoor criterion relative to (X, Y) and if P (x,z) > 0, then the causal effect of Y is identifiable and is given by the formula:"

$$P\big(Y = y \big| do(X = x)\big) = \sum_{z} \sum_{x'} P(Y = y |\, Z = z, X = x') P(X = x') P(Z = z | X = x)$$

In our case, the frontdoor path is the direct causal path X → Z6 → Y. By assuming that the variable Z6 is observed, the way of reasoning is intuitively as follows. Firstly, we can estimate the average causal effect of X on Z6, because there is no unblocked backdoor path between X and Z6. Secondly, the diagram allows us to estimate the average causal effect from Z6 to Y. There are some backdoor paths, like Z6← X← Z1→ Y. This is an open path that can be blocked by conditioning on X. Once we know the average causal effect of X on Z6 and the average causal effect of Z6 on Y, we can combine those to obtain the average causal effect of X on Y which gives us the formula of the backdoor criterion. In this way we can deconfound the causal relation between X and Y without using the unobserved Z1. This criterion is closely related to instrumental variables. It allows to control for confounders we cannot observe, including those we cannot even name.

$$\mathrm{X} \rightarrow \mathrm{Z6}: \; \mathrm{P}(\mathrm{Z6} = \mathrm{z6} \,|\, \mathrm{do}(\mathrm{X} = \mathrm{x})) = \; \mathrm{P}(\mathrm{Z6} = \mathrm{z6} \,|\, \mathrm{X} = \mathrm{x})$$

$$\mathrm{Z6} \rightarrow \mathrm{Y}: P(Y = y \,|\, \mathrm{do}\ (\mathrm{Z6} = \mathrm{z6})) = \sum_{x} \mathrm{P}\big(\mathrm{Y} = \mathrm{y} \,|\, \mathrm{Z} = \mathrm{z}, \mathrm{X} = \mathrm{x}\big)\mathrm{P}(\mathrm{X} = \mathrm{x})$$

$$\mathrm{X} \rightarrow \mathrm{Y}: P\left(Y = y \,|\, \mathrm{do}\ (\mathrm{X} = \mathrm{x})\right) = \sum_{z} P(\, Y = y \,|\, \mathrm{do}\ (\mathrm{Z6} = \mathrm{z6}))\ \mathrm{P}\left(\mathrm{Z6} = \mathrm{z6} \,|\, \mathrm{do}(\mathrm{X} = \mathrm{x})\right)$$

$$= \sum_{z} \sum_{x\prime} \mathrm{P}\big(\mathrm{Y} = \mathrm{y} \,|\, \mathrm{Z} = \mathrm{z}, \mathrm{X} = \mathrm{x}'\big)\mathrm{P}(\mathrm{X} = \mathrm{x}')\ \mathrm{P}(\mathrm{Z6} = \mathrm{z6} \,|\, \mathrm{X} = \mathrm{x})$$

2.4. Do-Calculus

The do-calculus is a generalization of the backdoor and frontdoor criteria (Pearl, 1995). It is used to find out if we can infer a causal relationship between 2 variables with observational data or if we need an experiment. In other words, it is used to assess if we can transform an expression containing the do operator into an expression without do-operator, which implies we can find causal relationships with observational data. The do-calculus contain 3 rules that should be repeated until the expression does not contain do-operators anymore.

- **Rule 1: Decide if we can ignore an observation**: any observational variable can be deleted from the expression if it does not influence the outcome through any path or if it is d-separated from the dependent variable.

$$P(Y \mid do(X), Z, W) = P(Y \mid do(X), Z) \qquad \text{if } (Y \perp\!\!\!\perp W \mid Z, X)$$

- **Rule 2: Decide if we can treat an intervention as an observation**: any intervention (do(x)) can be treated as observation (x) if a set of variables block all backdoor paths from X to Y. (derived from the backdoor criterion)

$$P(Y \mid do(X), Z) = P(Y \mid X, Z) \qquad \text{if } (Y \perp\!\!\!\perp X \mid Z)$$

- **Rule 3: Decide if we can ignore an intervention:** we can delete an intervention from the expression if there is no causal path (no unblocked path) from X to Y.

$$P\left(Y \mid do(X)\right) = P(Y) \qquad \text{if } (Y \perp\!\!\!\perp X)$$

The do-calculus has been found to be complete. This means that if a causal relationship can be found with observational data, then the do-calculus is able to transform an equation containing a do-operator, into a do-free equation. If do-calculus fails to do so, it means it is impossible to find the causal relationships with observational data and we should do an experiment (Huang, 2012).

An algorithm has been created to automate the steps of the do-calculus. The algorithm needs a causal diagram as input and is asked to find a do-free expression for a queried causal effect, if it exists. If the algorithm does not return a do-free expression, it is not possible to infer the causal relationship with observational data (Shiptser, 2006).

2.5. Discussion about the theory – or the battle between a famous computer scientist and a Nobel prize economist

Most of the theory described above has been developed by Judea Pearl, a computer scientist, who dedicated a big part of his research on causal inference with the use of causal diagrams. But there has been some discussion in economic literature about the relevance and usefulness of those graphs. Guido Imbens and Judea Pearl have had a scientific fight on whether causal diagrams are useful or not for economics. Pearl is promoting causal diagrams, Imbens finds common statistical tools used by econometricians more convenient. They have both written papers answering each other in recent years (Pearl, 1995; Imbens, 2014a; Imbens, 2014b; Pearl, 2014a; Pearl, 2018; Imbens, 2020; Pearl, 2020).

In 1995, with his article "Causal diagrams for empirical research", Pearl was one of the first ones to propose causal diagrams for empirical research. He introduced most of the theory seen in section 2 which shows that conditional independence can be extracted from the topology of graphs and that this is useful to tackle the confounding bias. Moreover, he introduced the do-calculus which can help causal scientists to know if it is possible to find a causal relationship thanks to graph topology. He highlighted other advantages of causal diagrams, which include: clarity of assumptions, transportability of knowledge, being able to do causal inference with a big number of variables, etc. He admits that the limitation of this theory is, that it reposes on the assumption that the graph represents the true relations and that absence of relations is difficult to prove. However, he claims that statistics is not enough to infer causality because it cannot distinguish between an observation and an experiment. He criticizes economists for not using this theory while other fields like epidemiology have embraced it. He questions:

> "*are problems in economics different from those in epidemiology? I have examined the structure of typical problems in the two fields, the number of variables involved, the types of data available, and the nature of the research questions. The problems are strikingly similar.*" (Pearl, 2014a)

Several authors discussed the paper of Pearl (1995), the statistician David Freedman argued:

> "*Pearl has developed mathematical language in which causal assumptions can be discussed. The* ***gain in clarity*** *is appreciable. The next step must be validation: to make real progress, those assumptions have to be tested.*" (Freedman, 1995)

D.R. Cox argues "*Graphical models and their consequences have* ***much to offer*** *here and we welcome Dr. Pearl's contribution on that account.*" (Cox, 1995).

On the other hand, Guido Imbens and Donald Rubin argued:

> "*We feel that Pearl's methods, although formidable tools for manipulating directed acyclical graphs, can easily* ***lull the researcher into a false sense of confidence*** *in the resulting causal conclusions. Consequently, until we see* ***convincing applications*** *of Pearl's approach to substantive questions, we remain somewhat* ***skeptical about its general applicability*** *as a conceptual framework for causal inference in practice*" (Imbens, 1995).

Recently, Imbens compared the use of the **potential outcome framework (PO)**, which is still the most famous framework in econometrics literature with the use of causal diagrams for causal inference (Imbens, 2020). Recent econometric textbooks discussing causal inference do **not contain causal diagrams** and focus on the potential outcome framework. In his paper, Imbens discusses the pros and cons of the 2 approaches for empirical work in economics trying to answer why most of the work in economics is closer to the PO spirit.

According to Imbens, there are 5 arguments that could be behind the **fame of the potential outcome framework** (which most economists are familiar with):

1. Some assumptions like monotonicity[1] (Angrist, 1994) or shape constraints like convexity or concavity (Matzkin, 1991; Chetverikov, 2018) are easy to highlight in the PO framework, this is not the case for causal diagrams.
2. The PO framework is easy to use in traditional approaches of economic modelling. For example, in demand and supply models, the potential outcome functions are natural primitives.
3. Many identifications of causal relationships in economics rely on the analysis of few variables where identifications strategies already have answers for.
4. The treatment effect heterogeneity is well considered in the PO framework. (Imbens, 1994; Sekhon, 2020)
5. The PO framework has long established the different inference strategies we need to identify a causal relationship.

Moreover, he argues that the theory **only works given that the causal diagram is correct** and that the causal diagram literature does not give real life examples of their use for empirical research, often focusing on "toy models" in the theory. He declares "***the most important issue holding back the DAGs is the lack of convincing empirical applications***." (Imbens, 2020). However, he acknowledges that causal diagrams are handy to illustrate the main assumptions of the model and that they can simplify causal inference in complex models with a lot of variables. He argues that the 2 methods (PO and DAGs) are **complementary** and that they have different weaknesses and strengths which make them both suitable depending on the question

[1] The monocity assumption implies that there are no defiers: assignment to the treatment group can only increase the probability of getting the treatment.

we want to answer, but that some of this is a matter of taste (given the same assumptions, some prefer algebraic representations, some prefer graphical ones).

Pearl counterclaims each of those arguments (Pearl, 2020) and argues there are way more empirical applications of the PO framework because in the 1990s economists were warned that graphical models were "confusing" and "deceptive" and that they had no scientific support.

> "*Not a single paper in the econometric literature has acknowledged the existence of structural causal models (SCM) as an alternative or complementary approach to PO*." (Pearl, 2020).

Pearl claims that this is why economists have not used this tool and that it appears in few empirical papers. On the other hand, in epidemiology causal diagrams became the traditional way of expressing assumptions (Pearl, 2020).

This debate does not seem to come to an end as both researchers keep their position defending proudly their tools to do causal inference. However, both authors have acknowledged that both tools are not completely useless. A hypothesis against the use of causal diagrams could be that they are time consuming to construct and that it is easier to perform statistical verification, in the place of reading dozens of articles to try to construct a causal diagram. What if a global causal diagram was openly accessible connecting all the variables studied in economics? If this existed, the researcher would just have to zoom into the relationship that he wants to study and pick an already made causal diagram. Maybe this would **enable causal diagrams to provide their full potential**. By making causal diagrams easier to construct and more accessible, we could give empirical proof of their adequacy or in contrary we could be able to prove without doubt that the potential outcome framework is indeed better suited for economics. In the next section we will examine the different methods used to build causal diagrams.

3. The different ways of building a causal diagram

Before being able to extract information from a diagram we have to construct one. The 2 main methods used nowadays are: making a diagram manually and making one with numeric data using an algorithm. Those diagrams are often made around one specific causal relation that we want to analyze. The third, new way of making a diagram and which we will try to implement in this thesis is the automatic extraction of a graph from scientific literature.

3.1. Manual causal diagram: done by researchers reading articles

Causal diagrams are usually made by scientists after having examined pertinent scientific literature on a specific topic. They represent the author's expert knowledge and understanding of a causal relationship among variables in a certain domain, they represent an image of how the author thinks the world works. This method is sometimes used to give more clarity to the assumptions in economic papers, but it is often not formally given to readers and stays a mental image. Figure 10 illustrates the different steps to make a causal diagram manually.

1. Reading 2. Thinking 3. Drawing

Figure 10: Illustration of a manual causal diagram

3.2. Data-based causal diagram: PC algorithm

Another way of constructing diagrams is with numeric data. In the last 20 years several algorithms have been developed to find conditional independency between variables. We can run conditional independence tests via partial correlations. One of the oldest one is the PC algorithm developed by Peter Spirtes and Clark Glymour (Spirtes, 2000).

Assumptions of the model:

1) The model is acyclic
2) Causal sufficiency: There are no hidden variables (no latent confounder)
3) Causal faithfulness: d-separation implies certain conditional independence relationships, but the other way around is not necessarily true. Causal faithfulness assumes that the reverse is indeed true (Heinze, 2018)

Given acyclicity, causal faithfulness and causal sufficiency we can apply the **PC algorithm to infer a causal diagram compatible with the data**. To apply the PC algorithm, we need a dataset containing all the needed variables. The algorithm finds all the conditional independence relations in the data and finally constructs a graph that is compatible with these conditional independencies. To illustrate how the PC algorithm works an example from (Glymour, 2019) is given. It is assumed that the true relationships are identical to Figure 11.

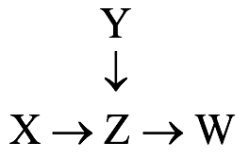


Figure 11: Example for the PC algorithm

From Figure 11 we can extract 3 paths: $X \rightarrow Z \leftarrow Y$ (collider), $X \rightarrow Z \rightarrow W$ (chain), $Y \rightarrow Z \rightarrow W$ (chain). The different steps are given to find the causal diagram presented in Figure 11, only by knowing conditional independencies between variables.

Step 1: Start with a complete undirected graph with edges bewteen all variables (Figure 12).

Step 2: Eliminate edges between variables that are unconditionally independent. Here X is independent of Y, so we eliminate the edge between X and Y.

Step 3: For each pair of variables connected by an edge (A-B). If there is a variable C connected to either of them (C-A-B or A-B-C), remove the edge between A and B if they are independent conditioning on C. Here X is independent of W conditioning on Z and Y is independent of W conditioning on Z. Therefore, the edge between X and W and the edge between Y and W are eliminated.

Step 4: For each triplet of variable (A, B, C) such that A and B are connected, B and C are connected, and A and C are not connected (A – B – C). Draw arrows $A \rightarrow B \leftarrow C$ (collider) if B was not in the conditioning set when the edge between A and C was eliminated. Here Z was not conditioned on when eliminating the edge between X and Y (step 2) so, $X – Z – Y = X \rightarrow Z \leftarrow Y$.

Step 5: We are left with 2 paths that are not completely directed ($Y \rightarrow Z – W$) and ($X \rightarrow Z – W$). We want to determine the direction arrow between Z and W. We have to choose between a chain ($Y \rightarrow Z \rightarrow W$) or a fork ($Y \leftarrow Z \rightarrow W$) (colliders have already been found in step 4). Because we know that the first edges in both paths are $Y \rightarrow Z$ and $X \rightarrow Z$ from step 4, we can deduce that $Z \rightarrow W$ and that the 2 undirected paths will form a chain. This is called the orientation propagation rule. In step 5 we end up with the same diagram than the one illustrated in Figure 11. There are several other orientation rules not illustrated here, but in some cases none of the orientation rules will apply to a given undirected edge and the edge will remain undirected as a result. This means that we will often finish with a **partially directed acyclic graph** that should be completed by intuition.

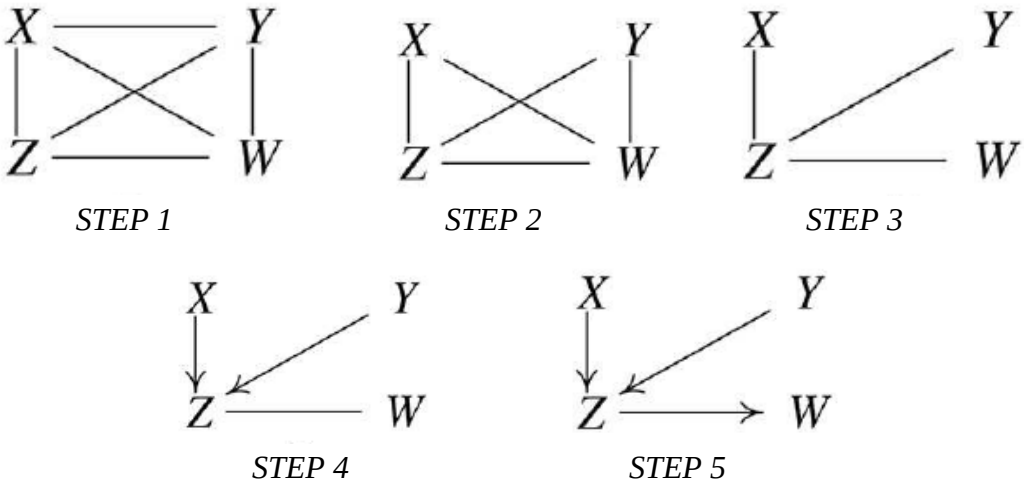


Figure 12: Illustration of steps followed by the PC algorithm (Glymour, 2019)

Since the PC algorithm, other algorithms which rely on different assumptions to extract diagrams from data were developed. The FCI algorithm (Fast Causal Inference algorithm) for example takes hidden variables into account (no causal sufficiency needed) (Spirtes, 2000). The comparison of the different existing algorithms can be found in (Glymour, 2019).

3.3. Automated literature-based causal diagram with machine learning

We have seen two of the most common ways to build causal diagrams in economics. First, a handmade diagram using the literature. Secondly an automatized data-based diagram using numeric data and an algorithm. The third way of making a diagram, which will be the focus of this thesis, is an **automated literature-based causal diagram**. It is based on analyzing literature using a trained supervised machine learning model to automatically extract relations between variables. This method is using Natural Language Processing which enables a computer to analyze human language and can be used to automatically extract information from unstructured text (see section 4.1). In order to be able to extract a causal diagram, a machine learning model can be trained on annotated data and can then be used on text it has never seen. The relations are extracted under the form of triplets (cause – relation – effect), those triplets can be used to construct a causal diagram. This method imitates the reading action of the scientist in an automatic way, like in section 3.1. It avoids forgetting variables and is less time consuming. It is linked to the second method because it is partially automatic and can handle a lot of data.

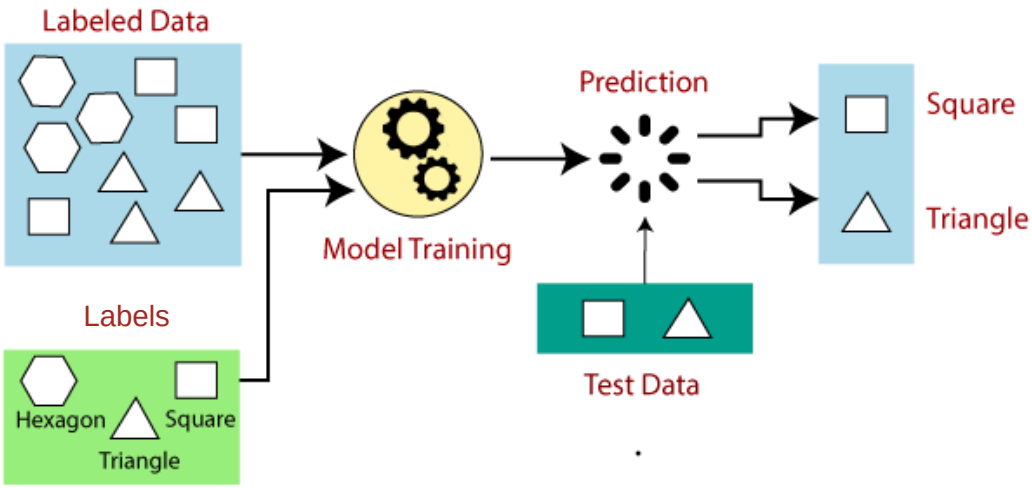


Figure 13: Illustration of a supervised machine learning model (javapoint, n.d.)

In Figure 13, the different steps involved in the creation of a supervised machine learning model are illustrated. At First, we need to define the different labels and annotate the training data with those labels. Subsequently, the model trains on the training data. Then, the model is able to extract labels from test data. Finally, the model compares the extracted labels with the manually annotated test set. By doing so, it determines its accuracy.

In the next section we will go through the evolution of information extraction and the different ways of doing it. Moreover, we will see that this method is used extensively in several domains like biology and epidemiology, but it is still underexploited in economics.

4. Existing literature using automated literature-based causal diagrams in economics and other fields

In recent years, computers became more and more capable of "**understanding**" natural language which made knowledge extraction more efficient (Chen, 2019). Knowledge graphs, which present information in a structured way, have become increasingly important. They are generated by automatic extraction of entities and relations. They are used mainly for question answering, decision making and predictions (Ashar, 2016). They are used by search engines and e-companies. For example, companies can use them to reference their products (Figure 14). Other examples of famous knowledge graph projects include DBPedia which extract entities and relations from Wikipedia pages (Lehmann, 2015) or Google Knowledge graph which is used by Google to enhance its search engine.

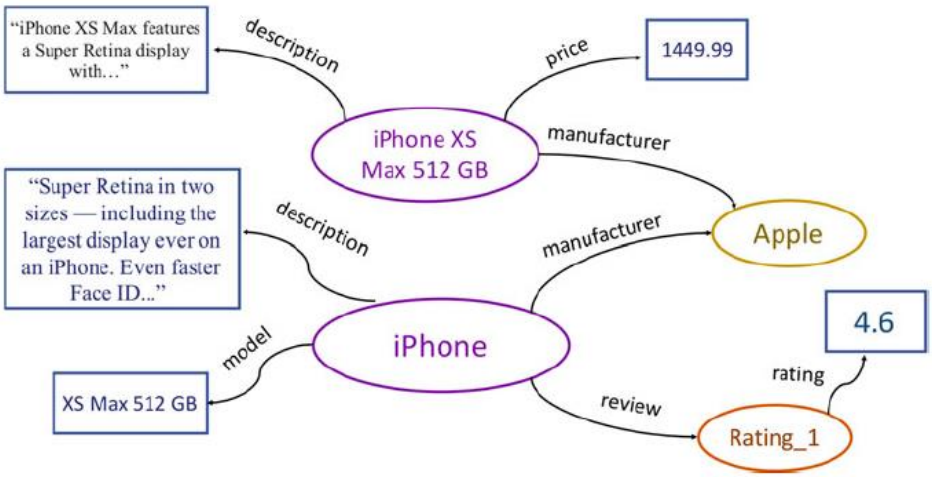


Figure 14: Illustration business knowledge graph (Kejriwal, 2019)

In the academic domain, knowledge graphs are mostly used in **bibliometric analysis** or citation analysis to link authors to their papers, publisher, years, organizations, research topics etc. (Figure 15). The bibliometric analysis is the main tool to classify articles and contains only few information. The only information about the content of the articles is in the title. Moreover, relations among researchers and research trends can be established. Some examples are Microsoft Academic Graph4 (Wang, 2020), Scholarlydata.org (Nuzzolese, 2016) and Open Academic Graph3. Those knowledge graphs are useful for researchers to make sense of research dynamics. But, huge parts of knowledge are neglected in those graphs because they **do not contain any information about the content of articles** and the manual extraction of this information is time consuming (Chen, 2019; Dessi, 2021).

> "*We still lack systems able to extract knowledge from large collection of research publications and automatically generate a comprehensive representation of research concepts. It follows that a significant open challenge in this domain regards the automatic generation of scientific knowledge graphs that contain an explicit representation of the knowledge presented in scientific publications.*" (Dessi, 2021).

In the scholarly domain, extraction of relations from papers where put into light by several SemEval[2] tasks. For example "*SemEval 2017 Task 10: Extracting Keyphrases and Relations from Scientific Publications" (Augenstein, 2017)* and *"SemEval 2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers challenge" (Gabor, 2018)*. Different methodologies of information extraction with the purpose to build graphs from articles spread in the literature (Li, 2019).

[2] SemEval is an international workshop on semantic evaluation whose mission is to advance the current state of art in NLP

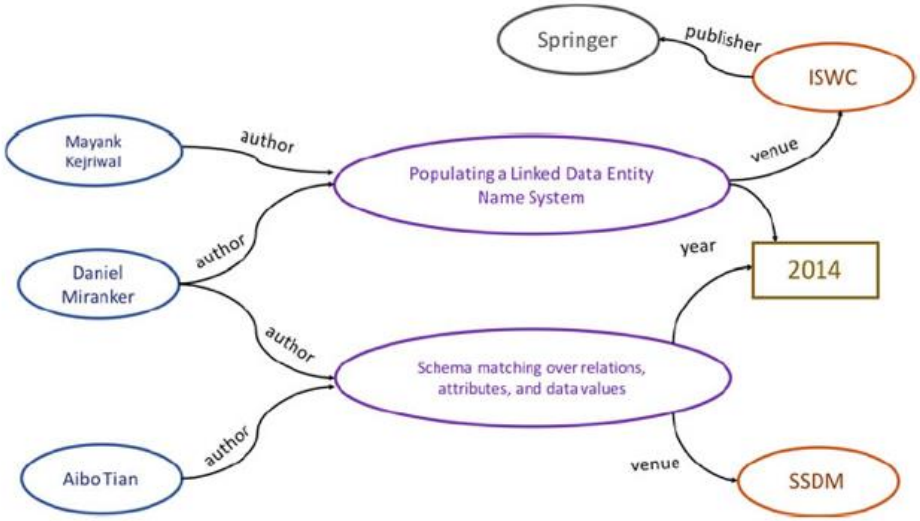
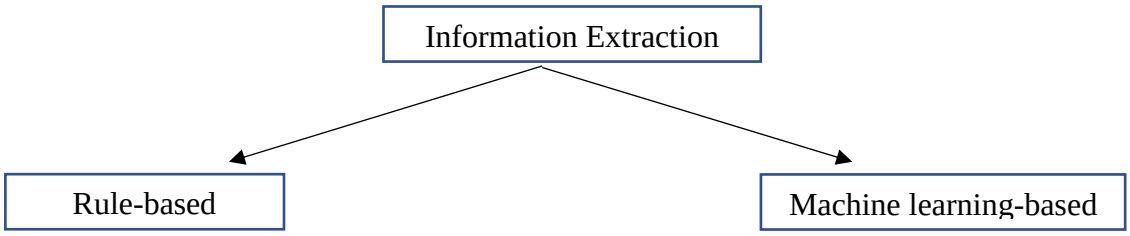


Figure 15: Illustration publication knowledge graph (Kejriwal, 2019)

For this thesis, the aim is to extract content from scientific papers and more specifically **causal relationships**. In section 4.1 we will discuss the literature of the different methods used for information extraction and more precisely causal extraction. Then in sections 4.2 and 4.3 we will discuss some applications of these methods in life-sciences and in economics.

4.1. Automated information extraction literature

In order to extract information from articles we need to use **Natural language processing (NLP)**. Natural language processing is an interdisciplinary field, that spans across several research areas such as artificial intelligence, computer science, linguistics and statistics. It enables computers to process large amounts of human language (unstructured data). The aim is to create a computer capable of "understanding" the meaning of text to extract information and insights like context, sentiment or writer's intent contained in the documents. Information extraction includes namely entity recognition and relation extraction (Yang, 2021) and can be split in 2 broad categories: rule-based methods and machine learning-based methods.



- *Rule-based models*

Until 1990, most NLP systems were rule-based which means that they included **handwritten rules** to find patterns. Information extraction was done by identifying patterns, keywords, sentence structure or syntactic analysis. Sentence structure relies on common patterns. For example, in English we often find the structure subject-verb-object.

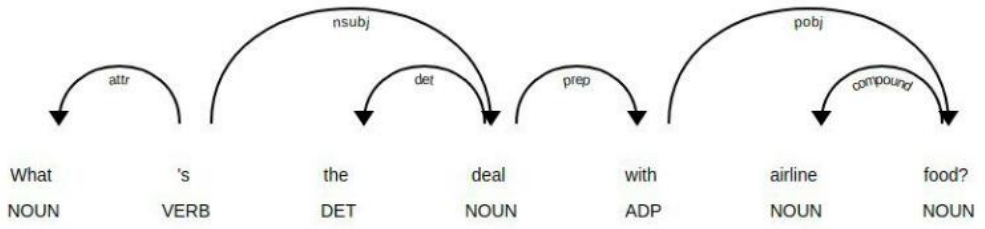


Figure 16 : Syntactic analysis (Mayo, 2018)

In Figure 16, the syntactic analysis is illustrated. However, there are an infinite number of rules that we can think of to extract information. For example, dates or names which can be listed to be extracted. This method is handy when the entities are present in a structured form in the text, or when the text includes explicit linguistic patterns (Tonin, 2017). Rule-based models are still broadly used in industries (Chiticariu, 2013). However, they **fail when there are many divergent ways to express the same thing**. Indeed, small differences in the formulation can prevent the model from finding the information that has to be extracted (Tonin, 2017).

- *Machine learning-based models*

From 1990, machine learning (ML) algorithms were introduced. In this case the statistical model is trained "***to learn the patterns of language from a set of examples***" (Fritzner 2017). Rules do not need to be explicitly given but a set of labeled documents is injected into the algorithm to train the model. The ML algorithm automatically focuses on the cases occurring the most and can be made more accurate by supplying more labelled/annotated data. This differs from rule-based systems which can only be made more accurate by increasing the number of rules or adapting them. Moreover, ML systems are usually more robust (Joachims, 2002) and attract more attention in academia.

There are 4 different kinds of learning models, including **supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning**. Supervised learning needs labelled data which is time consuming but is more accurate than unsupervised learning. On the other hand, unsupervised models do not need labelled data and search for common characteristics in the data, those models are less time consuming but also less accurate. Semi-supervised models use a mix of labelled and unlabeled data. Finally, reinforcement learning is based on the model taking decisions to optimize the reward. **Supervised learning is the most common method in the industry to analyze domain specific language** (Klein, 2005; Fritzner 2017). Furthermore, deep-learning is a subset of machine-learning and usually does not need human intervention.

	Pros	Cons
Rule-based	- Rules are explicit - High precision - Easy to adapt and add new rules - Easy to debug - Deterministic	- Cannot integrate the subtilities of natural language - Requires time to think of all the rules - Need often linguists
Machine learning-based	- Can be trained with examples - Can detect elements through subtle implicit rules - Probabilistic as opposed to deterministic - Easy to scale - Handle new data easily	- Often requires ML expertise - Rules are implicit so the models are hard to debug - Requires a lot of planning

Table 1: Pros and Cons of different methods

As we have seen, there are several methods to extract entities and relations from unstructured text. A subpart of the literature discusses the **extraction of causal relations** (Sorgente, 2013). There are many ways to express a causal relation, the simplest one is written in the form "*X causes Y*" or "*Y is caused by X*". These causal relations can be expressed by using many different types of propositions (e.g., subject-object, passive, active, nominal or verbal) and take many different syntactic forms (Ashar, 2016). The fact that causal relations can be expressed in many different ways also means that it is difficult to find patterns to extract them.

The expression of causal relationships is often classified in **explicit or implicit causality** and **intra- or inter-sentential causality**. Explicit causality expresses clearly the relation between entities and uses connectives that can be divided in the following categories. 1) causal links (e.g. because, therefore, hence, the result is, etc.) 2) causative verbs (e.g. contribute to, driven by, etc.) 3) resultative constructions 4) Conditionals (e.g. if…then) 5) Causation prepositions, adjectives and adverbs (Khoo, 2002). Implicit causality refers to relationships which are expressed with ambiguous connectives or sometimes no connectives (Yang, 2021). Intra-sentential causality describes a situation in which cause and effect find themselves in the same sentence. On the other hand, inter-sentential causality indicates that the relationship spans over several sentences. Table 2 illustrates an example of the different forms of causal relationships. Most studies about causal extraction focus on explicit causality with intra-sentential forms (Yang, 2021). However, in natural language causal relations can be expressed implicitly and/or in inter-sentential form, which is way more complicated to handle.

The forms of causal relations.

Sentences	Causality Forms	Causality Pairs
Financial stress is one of the main causes of divorce.	Explicit with Intra-sentential	<*Financial stress, divorce*>
Financial stress can speed divorce up.	Implicit	<*Financial stress, divorce*>
You may hear that unfaithful can lead to divorce. On the other hand, financial stress is another significant factor.	Inter-sentential	<*Financial stress, divorce*>

Table 2: Forms of causal relations (Yang, 2021)

Causal relationships can be expressed in many different ways, this is why they are difficult to formalize in a single grammatical model. Therefore, rule-based extraction models are not suited for causal extraction, it would be too labor-intensive and would not to extract implicit causal links (Ashar, 2016). Statistical-based models handle this better, and beside extracting explicit relations, they are also able to extract implicit relations (Yang, 2021). This is why a statistical machine learning model will be used to extract causal relations from economic literature (section 6).

4.2. Automated information extraction literature in life sciences

The use of causal diagrams to infer causality and the methods of automatic extraction of those diagrams have been exploited mostly in life sciences (Pearl 2014a; Pearl, 2018; Chen, 2020). For example, the biomedical field has a lot of knowledge bases like OpenPHACTS or the National Center for Biomedical Oncology BioPortal which are ontologies that represent genetic variations and their causal relationships with diseases[3]. In medicine causal relationships are important because they are used to determine if a medicine will cause the condition of a patient to improve or to understand which disease is the cause of certain symptoms. It has been shown that the automatic extraction of causal relationships is valuable in medicine and biology (Khoo, 2000; Sachs, 2005).

Some authors use abstracts from biomedical articles to identify the major findings and claims (Blake, 2010). Some use this technology to discover "new cancer driving mechanisms" (Valenzuela, 2018). Others use it to help researchers choose articles for making a literature review (Derchi, 2020).

[3] https://bioportal.bioontology.org/

Nordon et al. (2019) use this method to analyze medical records and construct a causal graph. Their result shows a greater precision in identifying confounding variables compared to medical domain experts, who have more difficulties to deal with a large number of covariates. Yu et al. (2019) showed that it is possible to identify accurately "*correlational, conditional causal, and direct causal statements*" from conclusions of medical articles. Some other interesting articles using information extraction technology in the medical sector are: Kitano, 2016; Lee, 2020; Bui, 2010; Mihaila, 2014.

Causality extraction has also been considered in media (Khoo, 1998; Balashankar, 2019).

We have seen that **automatic causal extraction is studied extensively in the domain of life sciences.** This is not the case in economic research. The next section will highlight the few papers discussing causality extraction in economic research.

4.3. Automated information extraction literature in Economics

Kim et al. (2021) **extract keywords from abstracts in economic literature** using deep learning. They argue that it is difficult to integrate existing natural language processing such as BERT[4] in economic literature because the language used in daily life differs from the language used in scientific papers. However, BERT has been used in several information extraction studies in life sciences.

Tilly et al (2021) show that we can extract knowledge graphs from economical newspapers to improve economic forecasting of industrial production.

H. Chen et al. (2019) extract information from **abstracts** about **construction management** using NLP. They divide the abstracts in four knowledge elements (background, objectives, solutions, and findings) to enhance bibliometric analysis. They propose an ontology to represent knowledge embedded in abstracts. The difference with this thesis is that they do not report the success rate of their model. Moreover, they extract connections between words but not causal relations.

V.Z. Chen et al. (2020) uses NLP to extract hypotheses from scientific papers in social sciences. This paper is to my knowledge the first example of **automated extraction of causal relationships** from economic literature (more specifically **business and management**). They focus on extraction of cause and effect because they are meaningful to highlight evidence. They argue that:

> "*in general, there is less clarity of the relationship between cause and effect entities in business articles relative to biomedical articles, thus making our task more difficult*". (Chen, 2020)

[4] BERT: Bidirectional Encoder Representation from Transformers. BERT is a language model developed by Google

First, they identify the hypotheses of the articles by extracting the text after "H" or "hypotheses" followed by a number. Once the hypotheses of several articles are extracted they extract causal relationships in those sentences, the accuracy of the model is 95%. This shows that **causal relationships can be extracted from literature** about organization performance. The difference with this thesis is that they do not differentiate between causation and correlation and consider causality to be causal or associative. This distinction will be made in the model created in section 6. Moreover, they extract the hypotheses of the articles, but they do not extract the findings. Furthermore, they only consider articles which include clearly stated hypotheses like "H1" or "hypothesis 1", we do not always find that in scientific articles, but this makes the model accurate since the sentences following H1 is often clear and short.
Izumi et al. (2019) use NLP on Japanese financial news articles and financial reports. They use supervised learning and more specifically a support vector machine (SVM) algorithm. In this thesis supervised learning is also used, the difference is that the algorithm used is based on a maximum entropy algorithm and that English text is examined. Moreover, this thesis focuses on extracting causal relationships from scientific papers which differs from news articles.
One of the closest work to this thesis was written by Yang et al. (Yang, 2020) who extracted links between economic variables from Chinese research reports and academic literature using a weakly supervised learning algorithm. In their paper, they construct a knowledge graph of the extracted information and use this graph as a tool for variable selection in economic forecasting. They argue that graphs made with NLP achieve a higher accuracy in economic forecasting in comparison with numerical statistical tools. They claim that with text mining, we can find economic variables that we wouldn't have thought of. They argue that Chinese reports mostly use narrative language which is easier to extract information from in comparison with English papers. Unfortunately, they do not give any information on the accuracy of their model (precision, recall, F1 score…), so it is difficult to assess if their model really works. The difference with this thesis is that we analyze English papers. Moreover, the links they used between variables are "increase", "decrease", "relate" which does not say much about causality.

All these papers were **written recently**. This shows that this topic has gained more attention in the last years but the number of articles focusing on causal relation extraction from economic literature is still low.

5. Motivation and potential application of an automated literature-based causal diagram

Some parts of this section have already been discussed in previous sections. However, this section summarizes the main motivations and potential applications in 7 points.

5.1. Growing number of papers / Time consuming

- Increasing scientific literature

Recent years have seen an **exponential increase of all kind of data**. This is also happening in the scientific literature. According to the National Science Organization report of 2017 (White, 2017), the academic literature doubled between 2004 and 2014. The Scopus database counts more than 1.4 billion references and 69 million articles and the ScienceDirect database counts over 250.000 articles which can be openly accessed (Chen, 2019). Scientific literature is considered as unstructured data often containing text in natural language, images and tables.

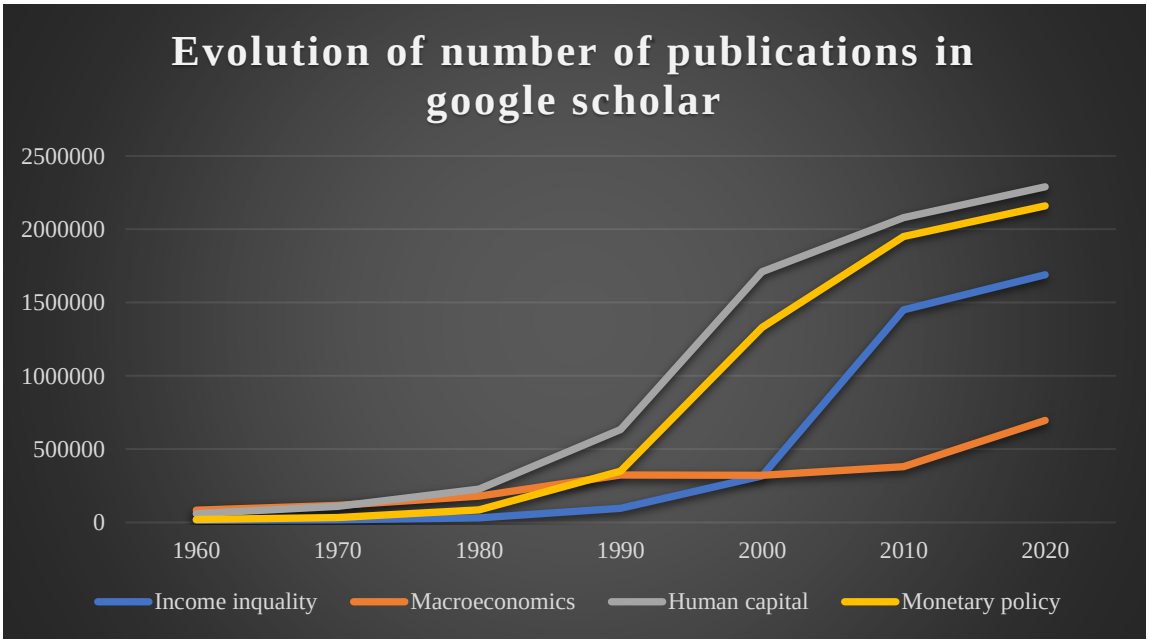


Figure 17: Number of publications in google scholar using keyword search

Figure 17 shows the increase of the total number of publications, resulting from a keyword search in Google scholar. 4 economical terms were used to assess the increase of publications: “Income inequality”, “Macroeconomics”, “Human capital” and “Monetary policy”. We can see that since 1990 the number of publications started to increase drastically. This is less the case for macroeconomics, maybe because this field has been studied more extensively in the past. However, for the other keywords the total number of publications in 2020 spans between 1.5 million and 2.5 million publications.

- Consequence: more time needed to read the literature

The main way to explore the existing literature is by manually reading through it. This is a **labor-intensive and time-consuming process** (Chen, 2019). Indeed, researchers spend most of their time on knowledge extraction and yet they are not able to read all the literature out there. Moreover, it became more and more difficult to make sense of all those publications. Usually, the researcher has to find a question, find who else is working on this topic and compare the different methods and data that have been used to answer this question. Reading and comparing hundreds of publications is an extremely cumbersome work while the number of publications is increasing every day.

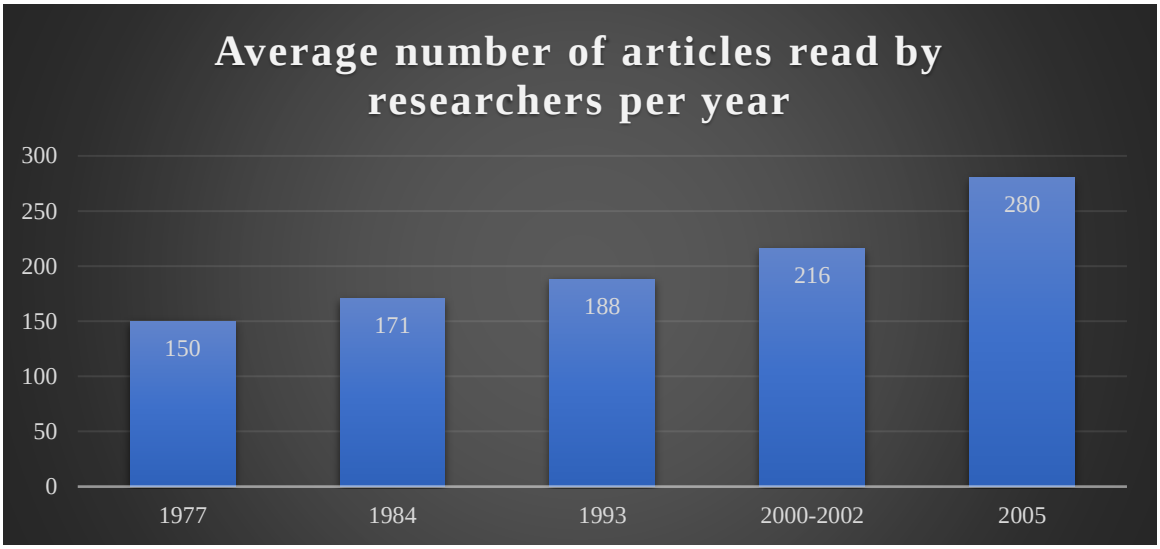


Figure 18: Numbers based on (Tenopir, 2009)

In Figure 18, we can see that the average number of articles read by researchers per year almost doubled from 1977 to 2005. This data comes from a survey made on U.S researchers in science faculties (Tenopir, 2009). Those numbers are pretty old, but we can assume that they continued to increase in the last 15 years. It has been estimated that researchers spend 23% of their total time reading articles (Hubbard, 2017).

Causal diagrams and more broadly knowledge graphs could help the researcher to have a better overview of the problem and consequently diminish the time spent reading. This would not make researchers obsolete and reading will still be part of their job. But it would accompany them in becoming more efficient (Auer, 2019).

Nevertheless, we have to make the automated causal diagram accurate enough, so that the results get close to the accuracy of manual work. Moreover, causal diagrams will only be time efficient for researchers if the training of the model does not take more time than actually reading the document. This is not difficult to assume if we consider that a model trained by one researcher could be reused by another researcher in the same field.

5.2. Recent discussion about Causal diagrams in economics

In section 2.5 we went through the debate about the use of causal diagrams for causal inference. Causal diagrams will not substitute the tools that have been used for years by econometricians (e.g. potential outcome framework). Each tool has its weaknesses and strengths, depending on what we want to analyze. For example, if we want to analyze a huge number of variables, there is no doubt that causal diagrams should be used. On the other hand, if we want to zoom into the relation between 2 variables, the potential outcome framework can integrate more specificities. However, this question is still debated. Making causal diagrams more accessible by having them automatically constructed will certainly increase the research around this topic and provide some answers.

5.3. Increased power of PC algorithm

In section 3.2 we have seen how the PC algorithm worked. This algorithm utilizes numeric data and is convenient to create a causal diagram automatically. However, with causal discovery algorithms we often finish with a partially directed acyclic graph because we cannot make the distinction between a chain and a fork. In future research, we could combine graphs made by numeric data and graphs made by text data to fix this problem. By doing this we could be able to direct the undirected arrows given by the PC algorithm. On the other hand, it would be interesting to study the differences of diagrams given by text data and graphs given by numeric data.

5.4. Causal diagrams graphs used extensively in other fields

In section 4 we have seen that knowledge graphs are extensively used in e-commerce and by search companies. Moreover, automatic causality extraction has been mostly researched in life sciences which have a lot of ontologies and knowledge bases (Auer, 2019). There are no clear reasons for causal diagrams to be underused in economics.

5.5. Variable selection and confounding bias

In section 2.3 we have seen that with causal diagrams, it is easy to fix the confounding bias and avoid open paths that create spurious correlations between the variables of interest. The confounding literature is closely related to the variable selection literature. Yang et al. use causal diagrams to select variables and perform economic forecasting (Yang, 2020).

5.6. Transparency, external validity and spread of knowledge

> "*One of the major accomplishments of causal diagrams is to make the assumptions **transparent** so that they can be discussed and debated by experts and policy makers.*" (Pearl, 2018).

Experiments are usually made in a certain population having specific characteristics. When we want to extrapolate results from an experiment to other populations we have to consider their differences in characteristics. **External validity** in economics refers to how effectively the outcome of a study can apply in other environments. Differences of environments and populations can be represented in a "selection diagram" where a selection node represents those differences. By using this diagram, it is possible to infer whether a causal effect in a target population can be deduced from an experimental finding in a different environment. Moreover, the results can be extended to observational studies, and we can transport outcomes from observational studies to other environments using "selection diagrams" (Pearl, 2011; Pearl, 2014b). External validity makes it possible to minimize measurement cost.

The spreading of scientific information is mainly based on unstructured data and analyzing it requires humans to read all those documents, this limits the **spread of knowledge**. Automatic processing of this knowledge will enormously increase the spread of knowledge (Dessi, 2021).

5.7. Artificial researcher

One of the ultimate goals of this line of research would be to create a global causal diagram in economics, which would represent all the knowledge from previous papers where each causal relation would be linked back to the papers that argue for this relation. This would allow us to immediately have the bigger view of the relation between variables, and to assess which relations do not have a consensus or have not been studied enough. Currently researchers ask questions based on their intuition. A **global causal diagram** could be used to automatically generate questions and hypotheses (Dessi, 2021; Auer, 2019). It would enable us to create an "**artificial researcher"** that would be able to help for decision making and question answering. It would also enable answering questions that we did not think of. Indeed, finding the good answer to questions is important, but finding good questions is also essential. Yang et al. claim that they are currently working on creating a website which would host a global diagram and encourage people to join (Yang, 2020).

We have gone through the theory of causal inference using causal diagrams, the different ways of constructing causal diagrams, the literature on automated literature-based diagrams and finally the motivations and potential applications. In the next section, a causal diagram will be built using economic literature. To do so, natural language processing and machine learning will be used to create a model of causal knowledge extraction. Secondly, the information extracted will be represented in a diagram in section 7.

6. Extract Domain specific information from text with IBM Watson Natural Language Understanding and Watson Knowledge Studio (WKS)

NLP is often difficult to implement for non-computer scientists. To tackle this problem IBM Watson proposes to make this kind of technologies accessible to everybody. IBM Watson commercializes software created for non-technical users to enhance performance of companies across the world. IBM Watson is the market leader when it comes to artificial intelligence and NLP (IDC, 2020). IBM Watson became famous in 2011 when it won the Jeopardy show quiz against finalists. It is now used worldwide across numerous industries.

IBM sells its software to companies and governments and provides free access to some of their tools to teachers and students under the "*IBM academic initiative program*"[5]. This program also provides training courses and teaching tools.

In this thesis we will use 2 services, namely Watson Natural Language Understanding (NLU) and Watson Knowledge Studio (WKS).

6.1. Watson Natural Language Understanding (NLU)

> *"IBM Watson Natural Language Understanding (NLU) uses deep learning to extract meaning and metadata from unstructured text data. It can extract categories, classification, entities, keywords, sentiment, emotion, relations, and syntax. Organizations can use these tools to interpret customer feedback, optimize marketing efforts, quickly understand current events, and analyze the latest market data at scale."* (NLU Presentation, n.d.)

A study from the independent consulting company Forrester, highlighted the economic benefices that NLU can imply for companies. The study claims that knowledge workers can **reduce by 50% the time** spend on text analysis (Forrester, 2021). Therefore, this tool can be interesting to apply to economic research.

To extract causal relationships from text, we need 2 features of NLU. Firstly, we need to extract **entities,** which are the **economic variables**. Secondly, we need to extract **relations** linking those economic variables.

NLU contains relation and entity types listed in the documentation (NLU Documentation, n.d.). It supports several languages including English. The NLU is already trained to recognize relations and entity types included in the service. The problem is that NLU only include **basic entity types** like: "Job Title", "Museum", "Website", "Person", etc.

[5] https://www.ibm.com/academic/home

Its entity types do not include an "economic variable" type which means that the NLU will not recognize that information. Moreover, NLU contains **basic relations types** like: "awarded to", "born in", "located at", "studied at", etc. but not "causal relations".

Illustration of how NLU works: If we feed the program with the sentence: "*Leonardo DiCaprio won Best Actor in a Leading Role for his performance*." and we ask NLU to extract the entities and relations, it recognizes that "Leonardo DiCaprio" is a "Person" and that "Best Actor" is an "EntertainmentAward" because, "Person" and "EntertainmentAward" is present in the type system of NLU. It also recognizes the relation "AwardedTo" between the 2 entities. If we do the same exercise with a simple economic sentence "*Inflation diminishes economic growth*", the NLU service does not recognize the entities "Inflation" and "Economic growth" nor the relation between the 2 entities, which is a causal relationship. This is problematic since this is a very basic sentence and that usually we can find way more complicated once. This implies that **NLU alone is not enough** to extract information from economic literature.

To cope with this problem some developers (Github, n.d.) use **rule-based models** that teach the service to recognize specific relations and entities that are not in the type system. We have seen that the problem with rule-based models is that it is very sensitive to different ways of expression one thing. Doing this for economic literature would be too cumbersome because there are many different ways to express a causal relationship in a language, with a lot of different words (impact, affect, increase, depends on, …). Moreover, giving the NLU service a list of entities for all the possible names of economic variables would also be too long to do because there are different ways to express the same economic variable.

To tackle the issue, IBM proposes to customize the NLU to the needs of your domain by using Watson Knowledge Studio. We will use Watson Knowledge Studio which enables to create a supervised machine learning model which has been highlighted in the literature as the preferred method for domain specific language.

6.2. Watson Knowledge Studio (WKS)

Watson Knowledge Studio (WKS) is a service that is part of IBM Watson, which proposes an AI service in the cloud. It is used to **teach IBM Watson the language of a specific domain** to make it capable of identifying entities and relationships from unstructured text in that domain.

This service can be used by domain experts **without writing any code.** It enables researchers in any domain to use artificial intelligence without the need for deep technical skills (WKS presentation, n.d.). No need to choose which kind of algorithm you will need, or to set all the parameters, this is already done by WKS.

Most of Watson products are available through an application programming interface (API), WKS however, has a **web interface**. On this web interface the whole process of model creation can be managed. A supervised machine learning model, a rule-based model or combination of both can be created. Once the model is created, it can be deployed to Watson Natural Language Understanding. On WKS one can upload documents and dictionaries, create a type system for the relations and entities, pre-annotate documents and manually annotate the documents with the relations and entities of interest. Once one has annotated enough documents, this training corpus can be used to train the machine learning model. The trained model is then used on the test set to check the model's accuracy. Once arrived at its best version, the model can be deployed to another service and be used to automatically extract information. The recommended practices for using this tool are described in the WKS documentation (WKS documentation, n.d.).

WKS exists since 2017 and has improved its features every year. Some features are still experimental like the cross-sentence relation recognition. IBM does not share which algorithm is used for its machine learning model, but Royan et al. (2020) argue that the algorithm is based on a **maximum entropy classifier**. A maximum entropy classifier is a probabilistic classifier which belongs to the class of exponential models.

6.3. Previous papers using WKS

To my knowledge, WKS has not been used to extract causal relationships between variables from economic literature yet. The service is quite new and mainly used commercially. It has been used in few scientific papers, mostly to extract information from medical reports. Tonin (2017) uses WKS to "*to extract mentions or indications of coronary artery disease in unstructured clinical reports*". He uses the machine learning model to find correlations between medication and coronary artery disease. Royan et al. (2020) used WKS for the DEFT challenge 2020 (DEFT, n.d.) which is a competition of information extraction. Their training corpus included 100 documents with on average 300 words. They focus on entity recognition in medical reports. The precision and recall of the model (F-measure 0,43 and 0,63) where very close to the results of other teams and show that **WKS is a competitive tool** in automatic information extraction. Derchi et al. (2020) argue that WKS can be used in assisting the researcher with literature reviews. They use 165 abstracts of restorative dentistry and annotate them. They aim to use Watson to identify the most relevant papers on a certain research domain. Fritzner (2017) uses WKS to analyze entities in emails in the shipping industry. He argues that WKS can reach larger target groups to take advantage of machine learning and claims that this method will have a key role in the future of NLP. Other scientific papers using WKS include (Georgescu, 2020; Singh, 2020; Laiq, 2020).

6.4. Create a machine learning model

The goal of the machine learning model is to **extract findings** from economic papers.

To train the machine learning model we need:
- documents to train the model
- a type system including relation and entity types
- a dictionary of economic variables (optional)

We will go through each step of the creation process namely:
1) the selection of documents
2) the creation of the type system including the entity and relation types
3) the pre-annotation
4) the annotation process and difficulties
5) the indicators of results
6) the results of the model, the different versions and finally its deployment

6.4.1. Documents

The documentation of WKS recommends to train the model on short texts that contains less than 2000 words. The supported file types for the documents are: CSV, TXT, PDF, DOC, DOCX, HTML, ZIP.

We are interested in extracting information from scientific literature in economics, but complete articles include too many words (50 pages on average for economic articles (Card, 2014)). They also include images and tables that are not readable by WKS. To tackle this problem, we could use conclusions or abstracts of articles. Indeed, both contain the major findings of articles. Chen (2019) argues: "*The paper's abstract, a concise and powerful statement describing the works in the paper, could provide more information about the paper.*" Moreover, they summarize the findings of the article in a short and clear way (between 100 and 250 words (Evans, 2021)). Abstracts are easy to access, because there is no need for the whole paper. Conventional journals are usually subscription-based, which makes only the abstract, references and citation information publicly available. There is no regulation in the way an abstract should be written. As a result, abstracts do not have a predefined structure. Moreover, journals do not require the abstract to have a specific content. However, it should offer the reader comprehensive overview of the paper (Chen, 2019). Conclusions on the contrary are often longer, are more difficult to access, and often include besides findings, recommendations for further work. Using conclusions would make the work of the machine learning model more difficult. This is why **abstracts will be used** to extract entities and their relations.

For the selection of the articles, one topic of economics was chosen to train the machine learning model. Only articles discussing "**income inequality**" in English were chosen. This was done to narrow down the vocabulary of the variable names. Indeed, it would take too much time to train a model that recognizes all possible variables and concepts in economics. Training a model often needs thousands of annotated documents to be performant. IBM recommends having minimum of 300.000 words in the model to yield sufficient results (WKS documentation). This is an issue because annotating the documents is time consuming. The amount of annotated data needed to obtain sufficient results differs depending on the size of the type system. Fortunately, the type system used in this setting is not long, neither complicated. However, the goal of this thesis is not to reach perfect precision of the model but to assess how it deals with economic language.

Literature reviews on income inequality were selected (Furceri, 2019; Bucevska, 2019; Hailemariam, 2018; Ichim, 2018; Hombres, 2012; Mdingi, 2021) and all the references present in those articles were listed. This gives a list of 277 different articles and books about income inequality. The selected articles are from different countries, and therefore writing style differences that could exist are addressed. Indeed, the use of causal language differs among authors from different countries (Yu, 2019). Secondly, we are only interested in the abstracts of the articles. Therefore, all the books and obsolete articles without an abstract were removed from the list. Moreover, only empirical articles were selected, because they highlight causal results in the abstract. Furthermore, articles that used unclear language were deleted: abstracts containing inter-sentential causal relationships or abstracts using implicit causality were ignored. After this, 102 abstracts were left and they contain each on average 120 words. These have been annotated manually and used to train the machine learning model.

The abstract of each article has been retrieved and separately stored in a DOCX file. Once all the abstracts were gathered, some of them were manually corrected. For example, some words were separated by a " – ".

6.4.2. Type system

The type system defines content that we want to label with annotation. It defines the types of entities and how relation between entities can be labeled. It was chosen after reading through some abstracts and was adapted later to suit better the annotation.

6.4.2.1. Entity types

3 entity types were chosen: variables, relation words and evidence. The **variables** are all the economic variables present in the text, the **relations** entities are the *causal connectives* words that relate the variables, for example: "affected by", "correlation between" or "impact on" etc. The **evidence** entities are the words, often in the beginning of the sentences such as "examine", "study", "results", "find".

This entity type is useful to identify if a relation is considered as a result or not. For example, in the sentence "*We study the relationship between inequality and growth*" the relation is not considered. On the other hand, in the sentence "*The results show a relationship between inequality and growth*" we take the relationship between inequality and growth into account because it is considered as a finding.

We talk about "economic variables" but since economies are extensively influenced by other things such as politics, health, etc., all the variables present in the article will be labeled as "economic variables".

6.4.2.2. Relation types

In the paper of Yang et al. (2020), 3 types of relations are used to relate economic variables: "increase", "decrease" and "neutral". They do not define exactly what they mean by "neutral". If we study the causal relationship between variables, we do not need to know if the relation between them is positive or negative to know if one variable is the cause of another variable. As seen in causal graph theory in section 2, the most important information we want to extract is if there is a causal relationship between those variables or not. Furthermore, some abstracts state that there is a causal relationship between 2 variables without saying if it is a positive or a negative relationship. For this reason, those relation types were not used for this model.

We will use the same relations Yu et al. (2019) used (except the conditional causality relation) to extract causal relationships: **causal relation, correlation or no relation**. Those are only 3 relations, which is good because the WKS documentation suggests limiting the number of relations and entity types.

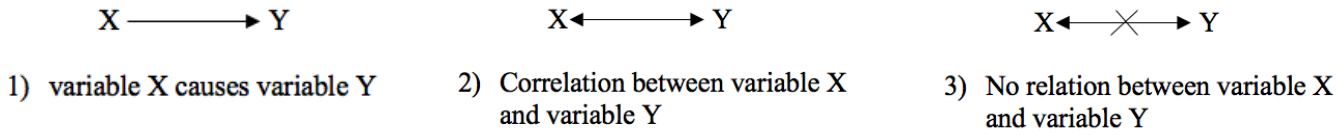


Figure 19: Relation types

Correlation is a statistical measure that describes the direction of a relationship between variables. A correlation between variables does not mean that one variable causes the other. **Causation** indicates that one event is the result of another event. This is also referred to as cause and effect. Finally, **no relation** indicated that there is neither causation or correlation relationship between variables.

6.4.3. Pre-annotation

Pre-annotation is optional, but it can be used to make the work of the annotator easier. Several annotation methods can be used in WKS.
Dictionary annotator: Adding a dictionary to the model can be used to pre-annotate documents. For example, if the word "inequality" is added in the dictionary under the entity type "variable" it will pre-annotate these words in all the documents without the help of a human annotator.

A list of economic variables used in the economic inequality literature is needed. The index of a handbook on inequality (Haugthon, 2009) is used to dress a list of 40 economic variables and their synonyms. For example: "globalization" = "economic globalization" or "demand for labor" = "labor demand" etc. The dictionary will not change the results of the model, it is just a helpful tool to reduce time consuming annotation.

Machine-learning annotator: This can be used once at least 15 texts have been annotated manually. A machine learning model can then be created and be used to pre-annotate the rest of the documents. The more annotated text, the better the machine learning model performs and the faster it goes to annotate new document because the pre-annotations made by the model have only to be corrected.
Rule-based annotator: The rule-based model was not used for pre-annotation because it cannot be used to label complex relations like causality (see section 4.1).
NLU-based annotator: The NLU-based model was not used to pre-annotate the abstracts because it does not recognize economic variables neither causal relations (see section 6.1).

Once the pre-annotation is done, all the pre-annotated documents were manually corrected to add the missing annotation.

6.4.4. Annotation Guidelines and difficulties

As previously mentioned, for this model a corpus of 102 abstracts was labelled. A given text can sometimes be annotated in different ways, depending on the interpretation of the type system and the language. Annotation requires the expertise of a domain expert who understands the meaning of the text. To train the machine learning model a **homogenous annotation** is needed because the machine will have difficulties to extract patterns from a heterogeneous data set. This means that mistakes in the annotation will give wrong examples to the model. This is why IBM recommends creating **annotation guidelines** if having more than one annotator. In the free version of WKS only one annotator account can be created. One person annotating the data yield a more homogeneous annotation than several annotators who could annotate slightly differently. The annotation guidelines were created and can be found in the Appendix.

The documentation advises to first annotate entities in the whole document, then the relations between those entities and finally the co-references. **Co-references** are used to identify multiple mentions of the same entity in a text. For example, in the text "*Inflation diminishes economic growth. It also has an impact on inequality*" we will co-reference the words "inflation" and "it" because it refers to the same thing in the text.

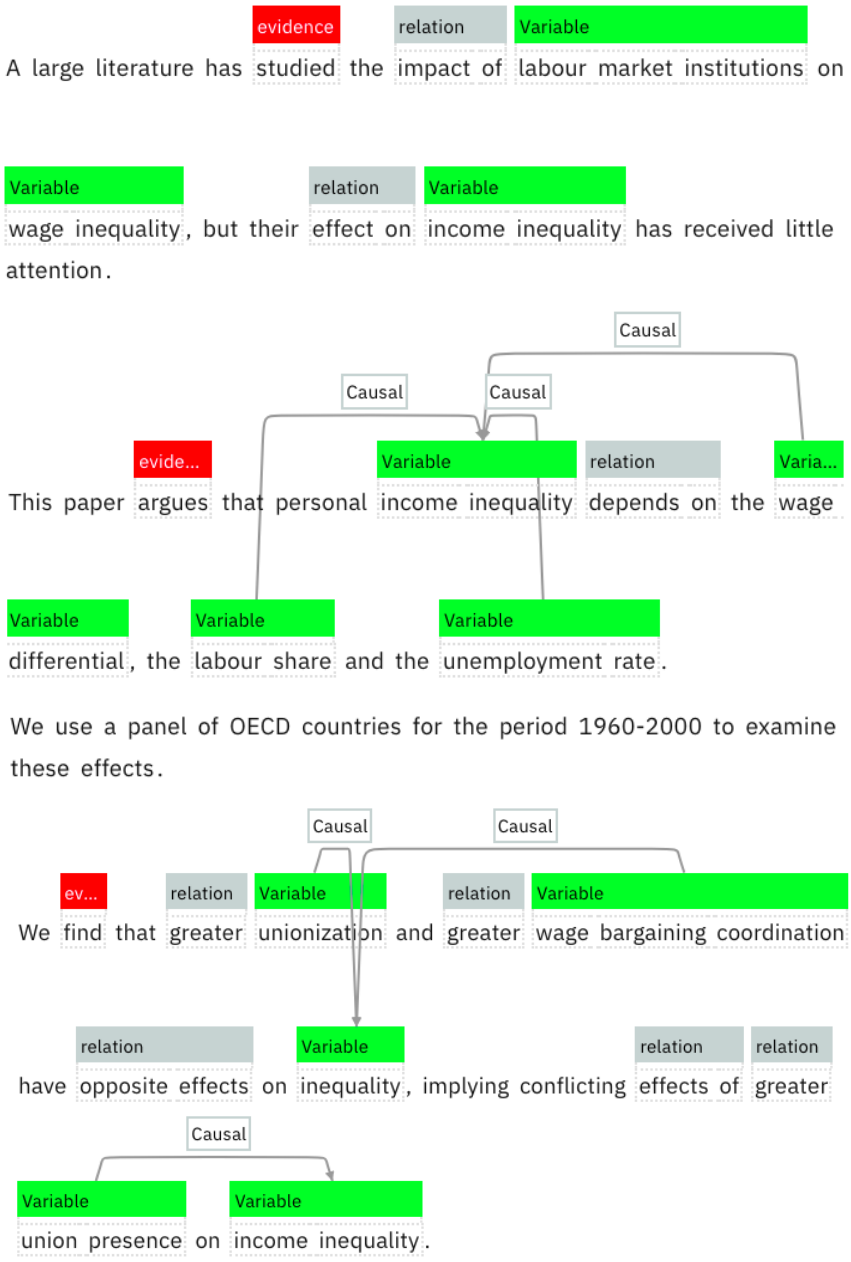


Figure 20: Example of an annotated abstract. Abstract from (Checchi, 2010)

In Figure 20, it can be seen in the first sentence that the relationship between "labour market institutions" and "wage inequality" is not considered because the evidence word "studied" is not an evidence word used for findings. On the other hand, in the next sentence the evidence word "argues" is present and the relationship is considered. Indeed, only the relations concerning findings are considered, which means that relations will only be considered when they are considered to be results (see annotation guidelines)

Several difficulties arose while annotating the documents:

- Dictionary problems

When different variables like “inequality” and “income inequality” which refer to 2 different things, are given in the dictionary, the pre-annotator will only annotate “inequality” in the documents. Moreover, when verbs are added in the dictionary it does not recognize the conjugated verb. This has to be fixed by the human annotator.

- Long economic variables

WKS documentation advises to annotate short passages, preferably 1 or 2 words. The problem is that economic variables often include more than 2 words. Some examples found in the abstract are: “income of the lowest decile of distribution”, “black women’s unemployment rate”, “married women’s labor supply”, “income share of the poor and the middle class”, etc.
In total 1341 variables were annotated in the different abstracts. The distribution of the number of words per variables is illustrated in Figure 21. The longest variable contains 12 words and the shortest once contain 1 word. In our 102 abstracts, 38% of the variables contained 1 word, 48% of the them contained 2 words, 8% had 3 words, 3% had 4 words, 3% had 5 words, the rest contained from 6 to 12 words occurring less than 3% of the time.

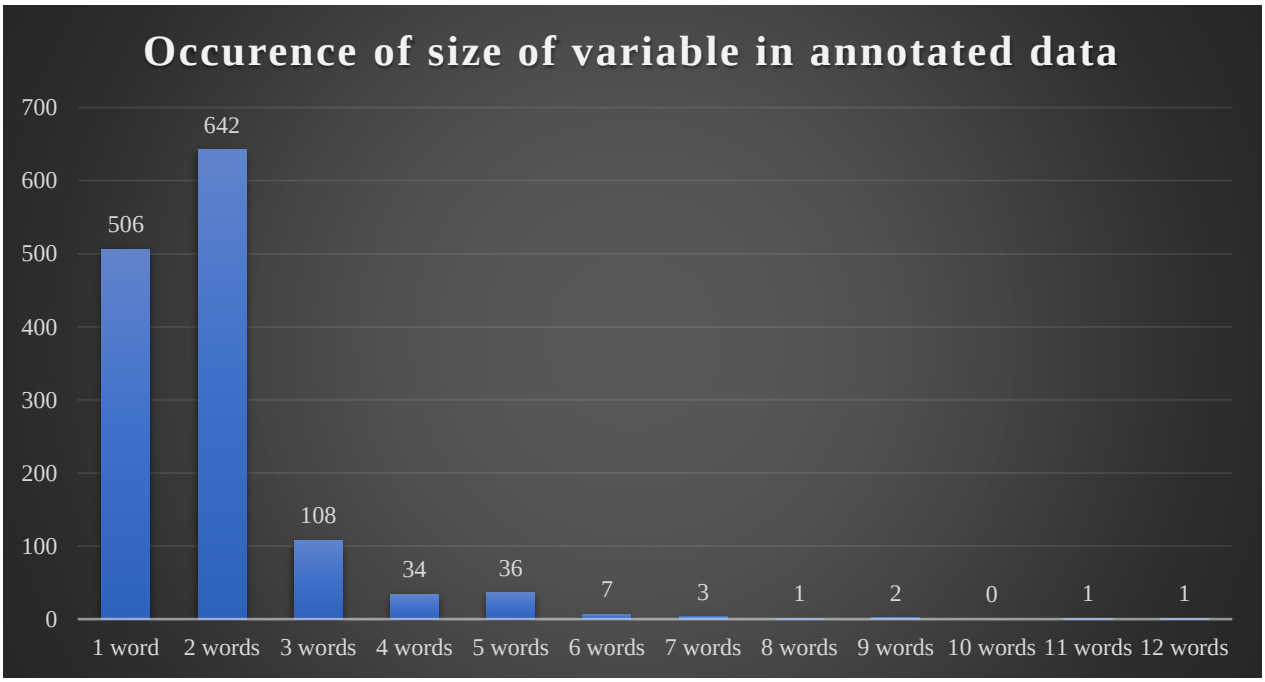


Figure 21: Number of words per variable

- Gathered variables

Some of the variables are put together in the narrative like “consumption and income inequality”, which imbricates 2 variables, “consumption inequality” and “income inequality” but this cannot be annotated in the text. When this happens, it has been decided to annotate the 2 variables together, because not annotating it at all could give wrong examples to the model. This is problematic because the documentation of WKS recommends to annotate not too many words together.

- Annotation of confounders

We have seen in the theory that confounders are very well represented by graphs. They have an important role in causal inference and not taking them into account could yield wrong conclusions.

An example from an abstract included in the training corpus: "*Both historical panel data and postwar cross sections indicate a significant and large negative relation between inequality and growth. This relation is only present in democracies.*" (Persson, 1994).

Using the type system, a correlation relation is labelled between "inequality" and "growth" (see Figure 22).

Figure 22: False causal diagram

The problem is that this annotation does not consider the fact that this causal relation is only present under some condition: being in a democracy. The correct causal diagram would be the one in Figure 23.

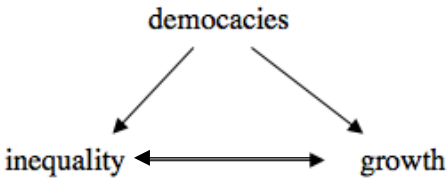


Figure 23: Real causal diagram

An entity type for confounders could be created, but that would make the identification way more difficult for the model. This is an issue that should be addressed in further work. For our model, confounders will not be included because it would make the identification of relations complicated. An idea could be to differentiate between "direct causality" and "conditional causality" like done in (Yu, 2019).

- Ambiguity of causal language

Sometimes the causal language in the literature can be ambiguous. We have seen in section 4.1 that causal relationships can be expressed in many different ways and can be implicit.
In the 102 abstract 71 of them expressed a causal relationship but very few of them used the word "causal". This is maybe because this word/concept has raised a lot of debate and authors want to be careful in the words that are used.
The annotation guidelines contain a list of "relation words" that are considered to highlight causal relationships (effect on, depend on, contribute to, etc.) and "relation words" who are considered to highlight correlation (e.g. relationship between, link between, correlation, etc.).
Some words can be used to express both kind of relation depending on the context. For example, in some cases the connection word "associated with" expresses correlation and sometimes it expresses causal link (which makes it more complicated for the machine learning model to find patterns).

Causation: “An increase in inflation is **associated with** an increase in income inequality”

Correlation: “A high inflation rate is **associated with** high income inequality”

In the example above, the first sentence denotes a causal relationship from inflation to income inequality. The second sentence denotes a correlation between inflation and income inequality. Although those 2 sentences are almost the same.

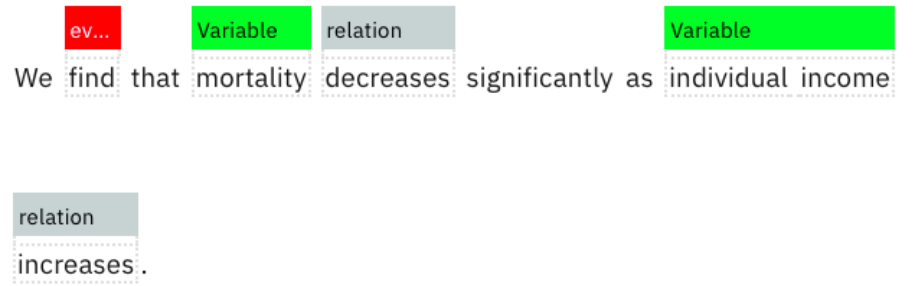


Figure 24: Abstract from (Gerdtham, 2004)

Sometimes it is even difficult for a human annotator to distinguish what the author meant. In Figure 24, it is not clear if we have a causal or correlation relation between mortality and individual income. When causation was not clearly expressed, the relation chosen was “correlation” because this relationship has less strong implications.

- Co-reference

Coreference’s cannot be annotated separately. In an abstract from (Haan, 2016): “*financial development, financial liberalization and banking crisis”* are denoted later as *“all financial variables*”. We cannot co-reference financial development with financial variables and then co-reference financial liberalization with financial variables. When this occurs, we do not co-reference those variables.

- Inter-sentential causality

Labelling cross-sentence, can be done in WKS but is still an experimental feature that can be discontinued. Moreover, it decreases the performance of the model. This is why causal relationships extending on more than one sentence were not annotated.

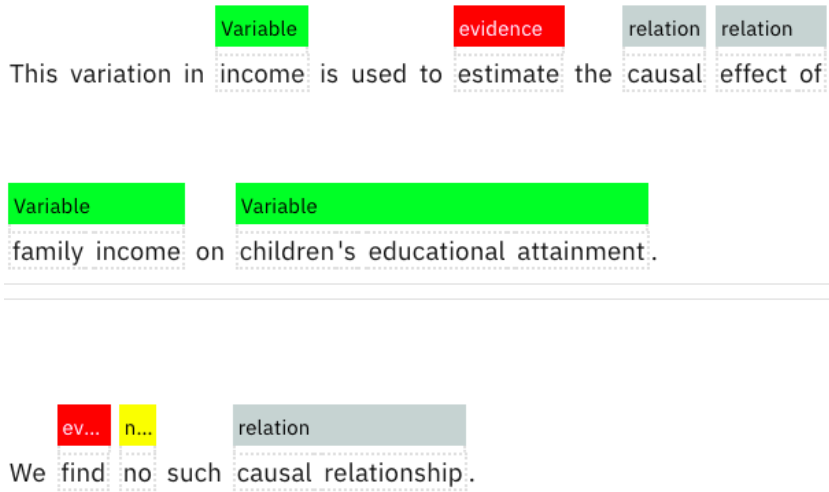


Figure 25: Abstract from (Loken, 2007)

Once all decisions were made concerning on how to perform the manual annotations and how to address the difficulties encountered. The complete document set was annotated.

6.4.5. Indicators of the results

Once a large enough set of documents were correctly annotated, the annotated documents were used to train the model and evaluate it.

The WKS documentation recommends using 70% of the documents to train the model, 23% of the documents to test the model and 7% for creating the blind set. Once the model is trained, it will annotate the test set and compare the answers given by the model to the correct annotations given by the human annotator. The test set is used to improve the model: by comparing the results of the model in the test set and the true results of the human annotated we can adapt the type system and the way we annotate. This is why a blind set is used to give final results. The answer given by the model in the blind set cannot be seen and the model can thus not be adjusted specifically to those examples. The detailed results of the blind set cannot be seen, because this could influence the way improvement are made.

- Several measures are used by WKS to determine the accuracy of the model

The **precision** score is the number of correctly detected annotations given by the machine learning model divided by the total number of labels given by the model. It measures the percentage of correct answers in comparison with the total answers given by the model. When we have a precision score equal to 1 for an entity or relation type, it means that all the entities or relations found by the model are correct. On the other hand, a low precision score means that the model labelled a lot of incorrect entities or relations.

$$\text{PRECISION} = \frac{\text{CORRECTLY DETECTED ANNOTATION}}{\text{CORRECTLY DETECTED ANNOTATION} + \text{UNCORRECTLY DETECTED ANNOTATION}}$$

The **recall** score is the number of correct annotations given by the model divided by the total number of annotations given by the human annotator. It measures the amount of correct annotations found by the model in comparison with the amount of annotations he should have found which is determined by the human annotation. When the recall score equals 1, it means that every entity or relation that should have been found, was found. On the other hand, a low recall score means that the model fails to find and label the correct entities and relations.

$$RECALL = \frac{CORRECTLY\ DETECTED\ ANNOTATIONS}{TOTAL\ NUMBER\ OF\ CORRECT\ ANNOTATIONS}$$

The **F1** score is a "*weighted average of the precision and recall values*" (WKS documentation).

$$\mathrm{F1} = \frac{2 * PRECISION * RECALL}{PRECISION + RECALL}$$

The **percentage of total annotations** shows how many words were labelled a given relation and entity type in comparison with the total number of words annotated.

The **percentage of corpus density** shows how many words labelled as a given relation and entity type in comparison with the total number of words (annotated or not). It can help to analyze how frequent some entity or relations types are in comparison with all the other words.

The **percentage of documents that contain this type** shows how many files include a given relation or entity type.

6.4.6. Results of the model/versions

The abstracts were uploaded, the type system was defined, dictionaries were added, and the documents were annotated. After that, the model was trained on the annotated documents and tested to assess the performance of the model. Different versions of the model were made trying to increase the precision and recall score following the recommendations present in the documentation. The free version of WKS allows to make maximum 10 different versions. The versions differ by the number of annotated documents, the number of entries in the dictionary, the type system and the consistency of annotations.

	Number of annotated abstracts	Entries in the dictionary	Type system	Comment
Version 1	26	Variables: 30 Relations: 10	Defined in section 6.4.2	
Version 2	22	Variables: 30 Relations: 10	Same as version 1	4 abstracts deleted from training
Version 3	22	Variables: 75 Relations: 45 Evidences: 21	Same as version 1	Increased dictionary entries
Version 4	22+ 23= 45	Variables: 75 Relations: 45 Evidences: 21	Same as version 1	Increased number of abstract annotated
Version 5	23	Variables: 75 Relations: 45 Evidences: 21	Same as version 1	Model trained by the last 23 abstracts only
Version 6	61	Variables: 75 Relations: 45 Evidences: 21	Same as version 1	Increased number of abstract annotated with ML model of version 5
Version 7	61	Variables: 75 Relations: 45 Evidences: 21	Entity type added for negation	Changed type system
Version 8	102	Variables: 75 Relations: 45 Evidences: 21	Same as version 7	Increased number of abstract annotated with ML model of version 7
Version 9	102	Variables: 75 Relations: 45 Evidences: 21	Same as version 7	Reviewed annotations for consistency

Table 3: Differences between the versions of the model

The best practice is to first annotate a small set of documents to define annotation guidelines and standardize the process. This is because the type system often needs to be adapted along the way and the annotations to be corrected (Fritzner, 2017).

	Entity Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
	Variable	0.55	0.55	0.55	64% (80/125)	12% (80/664)	100% (5/5)
	evidence	0.67	0.8	0.57	8% (10/125)	2% (10/664)	40% (2/5)
⚠	**relation**	**0.4**	**0.45**	**0.36**	**28% (35/125)**	**5% (35/664)**	**100% (5/5)**
	Overall Statistics	0.51	0.54	0.49	100% (125/125)	19% (125/664)	100% (5/5)

	Relation Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
⚠	**Causal**	**0.14**	**0.13**	**0.17**	**56% (25/44)**	**4% (25/664)**	**40% (2/5)**
⚠	**NoRelation**	**0**	**0**	**0**	**7% (3/44)**	**0% (3/664)**	**20% (1/5)**
	RelateCorrelation	0.57	1	0.4	36% (16/44)	2% (16/664)	60% (3/5)
	Overall Statistics	0.27	0.3	0.25	100% (44/44)	7% (44/664)	60% (3/5)

Table 4: Results test set of version 1

In **version 1** of the model, 26 abstracts were annotated. The dictionary contained 30 variables names and 10 relation names. Unsurprisingly, the first results are quite poor, because of the small training corpus. In Table 4 the entity types "variable" and "evidence" have respectively a F1 score of 0.55 and 0.67. The entity type "relation" has the lowest score (0.4). This is certainly because the training corpus is not homogenously annotated. For the relation types: "causal" relation has a low F1 score (0.14) while "correlation" has a higher F1 score (0.57). This could be due to the fact that there are less ways to express "correlation" in comparison to "causation", correlation is often expressed by "there is a correlation between" while for a causal relationship the word "causal" is almost never used. On the other hand, "NoRelation" has a F1 score of 0, this may be due to the low density of "NoRelation". All documents contain "variable" and "relation" entities while 40% of the documents only contain "evidence" entities. Moreover, 60% of the documents contain correlations while 40% contain causation and only 20% contain no relation.

The WKS documentation provides **recommendations to improve the results of the model**:

- Add dictionaries or add entries in the existing dictionary
- Increase the training corpus
- Add type specific documents
- Enhance human annotator guidelines
- Update type system

In **version 2**, 4 abstracts were deleted from the model because the language used was ambiguous to highlight relations. This increased the performance of the model for the entities the F1 score increased from 0.51 to 0.67, for the relation the F1 score increased from 0.27 to 0.33 (see Figure 26). More detailed results of the different versions can be found in the appendix. In **version 3**, annotation guidelines were created after going through the annotated documents to ensure the consistency of future annotations. Moreover, the annotated documents of version 2 were corrected following the annotation guidelines to increase the homogeneity of the training corpus. Furthermore, all the annotated entities in those documents were added to the dictionary to make the pre-annotation of the next documents easier. This gives 71 variable names, 45 relation and 21 evidence words in the dictionary. It can be seen in Figure 26 that making the annotation more homogeneous, increased the F1 score of both entities and relations.

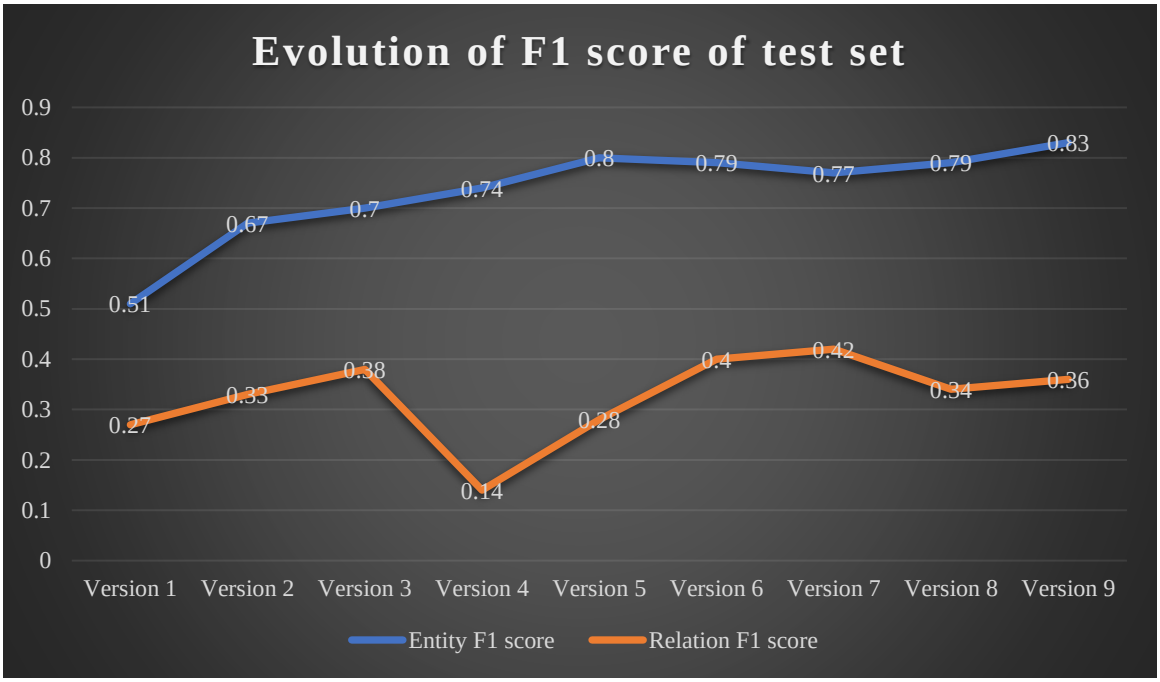


Figure 26: Evolution of results of the different versions

In **version 4**, 23 additional abstracts were annotated to get to a total of 45 annotated documents. By doing so the F1 score of the entities increased to 0.74. Surprisingly, the F1 score of the relations dropped significantly from 0.38 to 0.14. Increasing the number of annotated documents should increase the precision and recall of the model. It could be that the second set of documents has more implicit causal relationships. To understand those results **version 5** was created. In this version the model was only trained using the 23 additional documents added in version 4. By doing so, the F1 score went up again. So, when the model is trained with the 22 first abstracts and 23 last abstracts separately it has good results, but when the 2 together (45 abstracts) are used for training, it yields worse results. One explanation for this could be that the set of documents present an inhomogeneous language style. In **version 6**, 16 abstracts were added to the 45 already annotated. The machine learning model of version 5 was used to pre-annotate the new documents, which made annotation easier. In comparison with version 4 that used 45 documents to train the model, the F1 scores of version 6 are higher. For the entities the F1 score raised from 0.74 to 0.79, for the relation F1 score raised from 0.14 to 0.4. This increase is mainly due to the increase in F1 score of "causal" and "correlation" relations, respectively 0.39 and 0.63. However, the F1 score for the relation "NoRelation" is still zero, this is mainly due to the few examples of no relation in the annotated data which is present in 21% of the annotated documents in as opposed to 71% for causal relationships and 43% for correlation. This could be due to the **publication bias** that occurs in published academic research. This occurs when the result of a paper influences the decision whether to publish it or not (Song, 2010). Indeed, papers arguing that there is no relation between 2 variables are rarer, this could also be due to the fact that it is more difficult to prove there is no relation than to prove there is a relation (Imbens, 2020)(it is more difficult to prove that something does not exist than to prove that something exists). To try to make it clearer for the model to recognize "NoRelation", in **version 7** the entity type "negation" was added to the type system, and all the negations were annotated in all the previously annotated documents (e.g. no, not statistically…).

We can see in Table 5 that adding this entity type increase the F1 score of "NoRelation" by 0.2. On the other hand, it reduces the overall F1 score of the entities.

	Entity Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
	Variable	0.79	0.82	0.76	65% (292/447)	16% (292/1803)	100% (14/14)
	evidence	0.78	0.88	0.7	10% (43/447)	2% (43/1803)	93% (13/14)
⚠	**negation**	**0**	**0**	**0**	**2% (8/447)**	**0% (8/1803)**	**21% (3/14)**
	relation	0.79	0.88	0.72	23% (104/447)	6% (104/1803)	100% (14/14)
	Overall Statistics	0.77	0.83	0.72	100% (447/447)	25% (447/1803)	100% (14/14)
	Relation Types	**F1**	**Precision**	**Recall**	**% of Total Annotations**	**% of Corpus Density (by number of words)**	**% of Documents that Contain This Type**
⚠	**Causal**	**0.39**	**0.35**	**0.45**	**61% (104/170)**	**6% (104/1803)**	**71% (10/14)**
⚠	**NoRelation**	**0.2**	**0.5**	**0.13**	**18% (31/170)**	**2% (31/1803)**	**21% (3/14)**
	RelateCorrelation	0.63	0.75	0.55	21% (35/170)	2% (35/1803)	43% (6/14)
	Overall Statistics	0.42	0.43	0.42	100% (170/170)	9% (170/1803)	100% (14/14)

Table 5: Results test set of version 7

In **version 8**, 41 abstracts more were annotated, which gives a total of 102 abstracts. Those abstracts were pre-annotated by the ML model of version 7. This gives a total of 102 labeled abstracts. For the entity type recognition, version 8 performs better. This is mainly due to the increase of the F1 score for the negation entity that goes from 0 in version 7 to 0.27 in version 8. On the other hand, the results for the relation types are deceiving because the model of version 8 performs worse than in version 7 even though more documents where annotated. Indeed, the F1 score for "RelateCorrelation" goes from 0.63 in version 7 to 0.4 in version 8 and the "NoRelation" relation drops to 0.

Entity Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
Variable	0.81	0.86	0.76	65% (593/914)	19% (593/3153)	100% (23/23)
evidence	0.84	0.94	0.77	11% (99/914)	3% (99/3153)	100% (23/23)
negation	0.57	0.8	0.44	1% (9/914)	0% (9/3153)	26% (6/23)
relation	0.86	0.95	0.79	23% (213/914)	7% (213/3153)	100% (23/23)
Overall Statistics	0.83	0.89	0.77	100% (914/914)	28% (914/3153)	100% (23/23)

Relation Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
Causal	0.38	0.41	0.36	66% (225/339)	7% (225/3153)	78% (18/23)
NoRelation	0	0	0	10% (35/339)	1% (35/3153)	22% (5/23)
RelateCorrelat...	0.44	0.45	0.43	23% (79/339)	3% (79/3153)	48% (11/23)
Overall Statistics	0.36	0.39	0.33	100% (339/339)	11% (339/3153)	100% (23/23)

Table 6: Results test set of version 9

In **version 9,** the 102 previous annotations were reviewed to increase the homogeneity. The test set was analyzed to compare the results of the model with the human annotation. Often the model does not find the relationships because it fails to identify correctly the entities in the first place. In Figure 27 we can see that the model fails to recognize the entities “stock returns”, “household debt” and “rates”. The result is that it does not find the relationships including those entities. Another common cause of failure is that the model often mixes causal relation and correlation. This is not a surprise since it is sometimes difficult to distinguish, even for a human. Finally, another cause of error is due to relations that are annotated by the model and that do not represent findings. In Figure 28 we can see that the model finds a relation between variables while it is not a result. The results of version 9 show an average F1 score of 0.83 for entities and 0.36 for relations (see Table 6). For both the entities and relations, the precision is higher than the recall which means that the model does not make too many mistakes but that it does not find some entities or relations that should be found.

The **blind set gives comparable results** (see graphs of results in the appendix). It reaches an F1 score of 0.56 for entities and 0.05 for relations in version 1. The relation score is way lower in the blind set than in the test set in version 1. In version 9 the blind test gives an F1 score of 0.82 for entity recognition (with a precision of 0.87 and a recall score of 0.77). The result for the relation recognition is 0.33 (with a precision of 0.36 and a recall score of 0.77). Those results are similar to the test set, which means that the test set was a good representation of the total set of documents and that the adaptations made to the model did not only improve the efficiency of the model on the test set but also on the blind set.

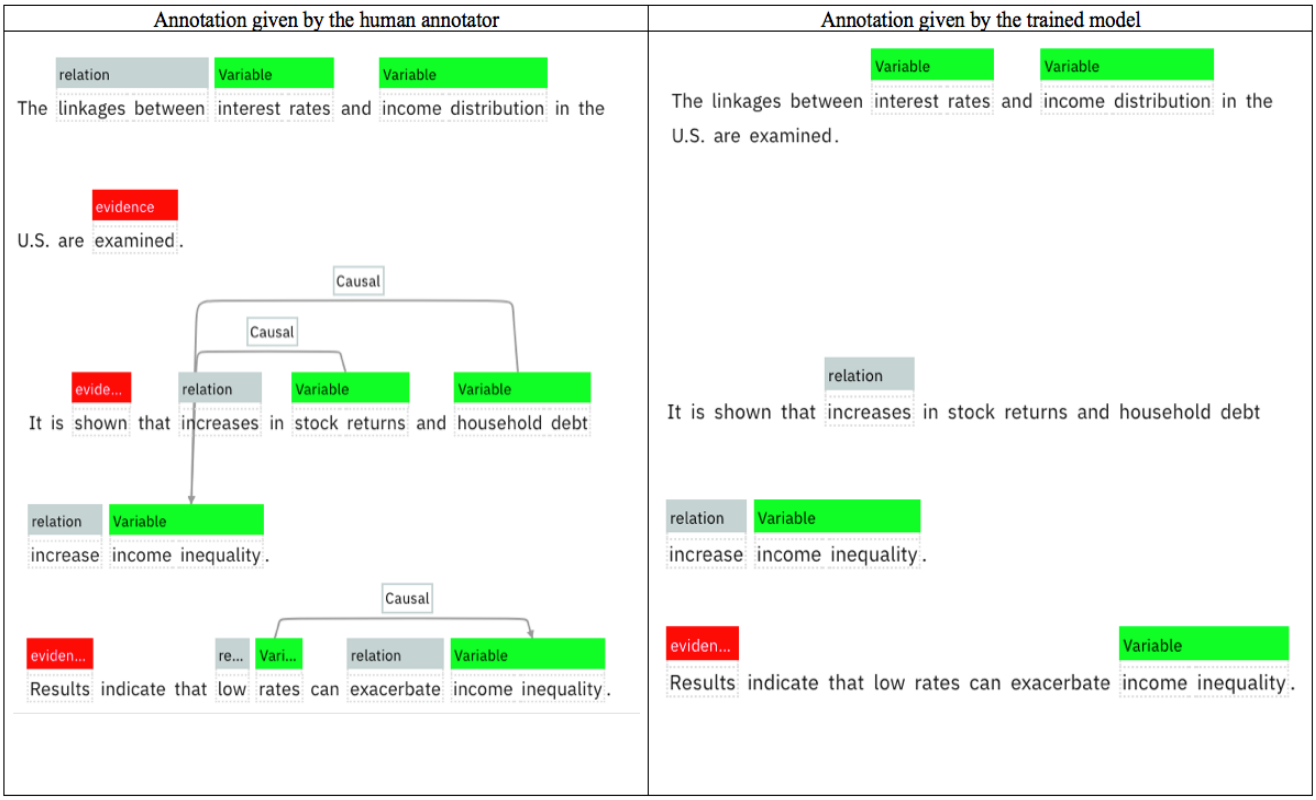


Figure 27: Example of relations that are not considered because entities are not found in the first place. Abstract from (Berisha, 2018)

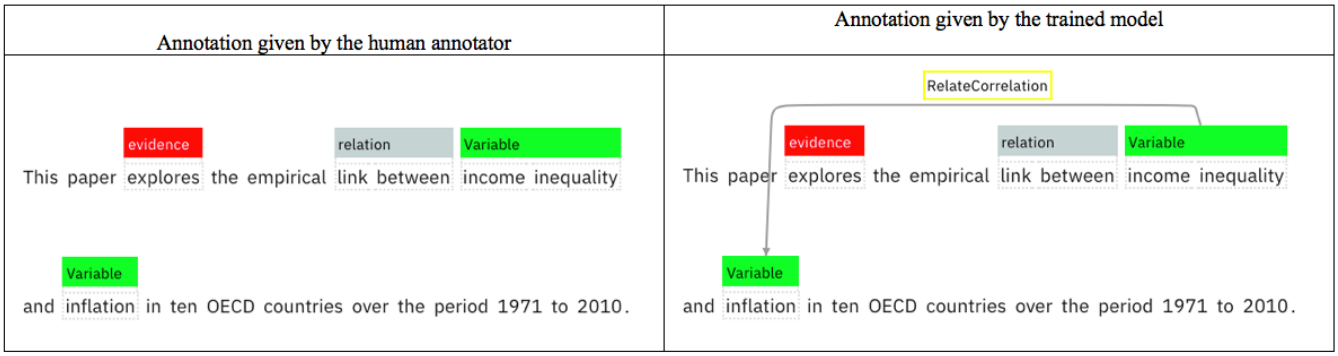


Figure 28: Example of relations annotated by the model that should not have been annotated. Abstract from (Monnin, 2014)

- Discussion on the improvement of the results

The best versions of the model are version 9 for entity recognition (F1 score 0.83 in test set) and version 7 for relation recognition (F1 score 0.42 in test set). Anyway, version 9 is considered as the best version because it has the best results for entity and relation recognition in the blind set (see graphs of results in the appendix). This model has a **high F1 score for entity** recognition, 0.83, which is close to other papers using WKS for entity recognition. On the other hand, the **F1 score for the relation types** has increased from 0.27 in version 1 to 0.36 in version 9, but this is still **low**. Izumi et al. (2019) for example found an F1 score of 0.81 from extracting relations from newspapers, this is maybe due to the fact that newspapers use a more narrative language in comparison with scientific papers.

The low F1 score for relations extraction is not surprising considering the **small amount of training data** with in total around 13.000 words in the training set while the documentation of WKS recommends having 300.000 words to yield sufficient results. Considering that the recommended amount of labelled data is 20 times larger than in our case, it can be argued that increasing the training data will increase the F1 score. Another way of explaining the low average F1 score for the relations, is the F1 score of "NoRelation" which equals 0. The results could be improved by increasing the examples given to the model for this kind of relations.
A copy of the annotated corpus, type system and dictionary can be found in a Github repository[6]. This data is open access and can be used in WKS. Moreover, version 9 of the model was deployed to the Natural Language Understanding service of IBM to create a causal diagram (see section 7).

6.5. Advantages and disadvantages of WKS on economic literature

PROS	CONS
- easy to use and intuitive for non-computer scientist - easy annotation on the web interface - free version for academic research - model can be deployed to different services of IBM - gives a probability of the relation to be correct (see section 7)	- less manipulatable than code - no precise performance analysis (e.g. how many relations found by sentence or average amount of words per entity)

Table 7: advantages and disadvantages of WKS

7. Creation of a causal diagram with Watson Natural Language Understanding (NLU) service

In order to find out if the model works, the results of NLU with the model created in section 6 and without the model are compared. In section 6.1 we have seen that the NLU service couldn't extract relation or entities from the sentence "*Inflation diminish economic growth*", when our customized model is added to NLU, the service finds 2 entities, "inflation" and "economic growth", that it recognizes as variables, and it finds that inflation has a causal relationship with economic growth.

[6] https://github.com/valeriedevos/IBM-Watson-knowledge-studio-economic-litterature/tree/main

Moreover, when we ask NLU without the model to extract entities and relations from the sentence "Rising trade and exchange rate crisis reduces income inequality." It does not extract any entity neither relation. Once the machine learning model is used to analyze the same sentence, it recognizes the entities "trade", "exchange rate" and "income inequality" as variables, "reduces" as a relation and finds 2 causal relationships: trade – causes – income inequality and exchange rate – causes – income inequality. The results are given in JSON format. The entities and relations found by the model are given with a **probability score** for them being correct. This is useful because if we want to analyze more articles we can delete the relations and entities with lower scores.

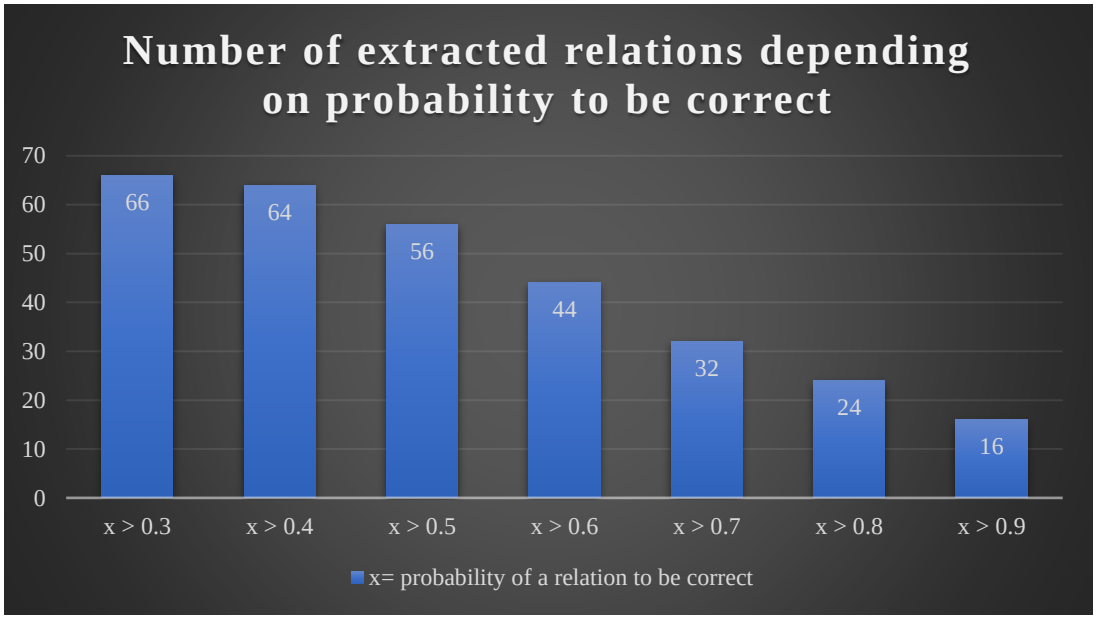


Figure 29: Number of relations by probability to be correct

We already know the performance of our model, but to show what it can do and construct a causal diagram, we use all the abstracts included in the test set, which were not used by the model for the training. We use the latest version of our model to extract the relations of those 23 abstracts. In total, the model finds 76 relations, after deleting doubles we get 66 unique relations. In Figure 29 we can see that on the 66 extracted relations, 56 of them have a probability higher than 50% to be correct and 16 of them have a probability higher than 90% to be correct.

From the JSON file the relations that have a higher score than 0.5 are extracted and stored in the form of triplets (cause – relation type – effect). A graph is then produced that links all the triplets together in a directed graph (directed from cause to effect).

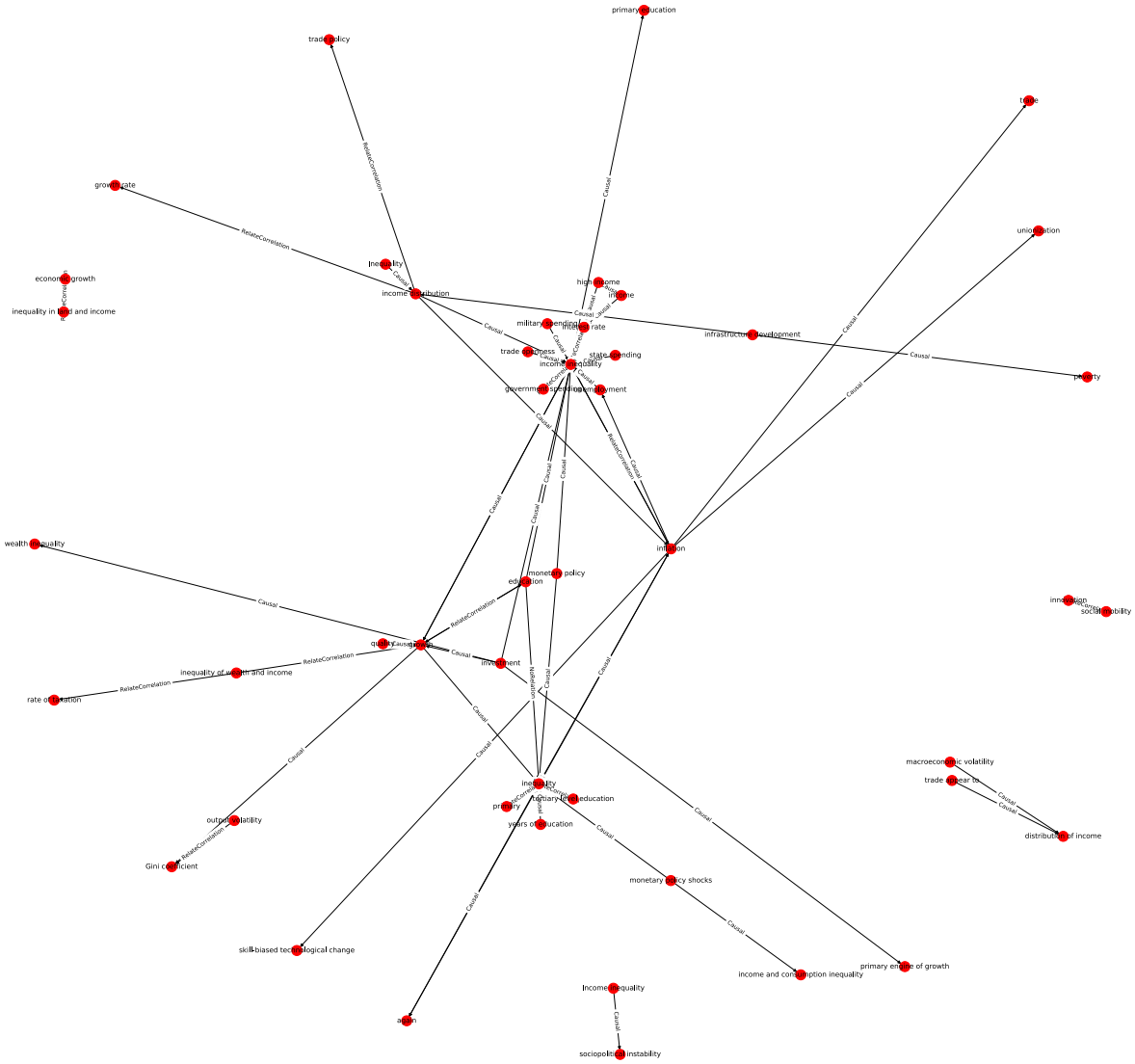


Figure 30: Causal diagram extracted from the test set

In Figure 30, 56 relations are represented in a causal diagram. The nodes that are the most connected are “Income inequality” and “inequality”. This is because most of the articles discuss this topic. Any other topic in economics could be imagined if the model was trained on other specific domains. Figure 31 is a zoom of Figure 30. Most of the relations are causal relations, followed by correlation and no relation.

A causal diagram was also made from the 102 abstracts to represent what a complete causal diagram about income inequality could look like (see Appendix 11.4).

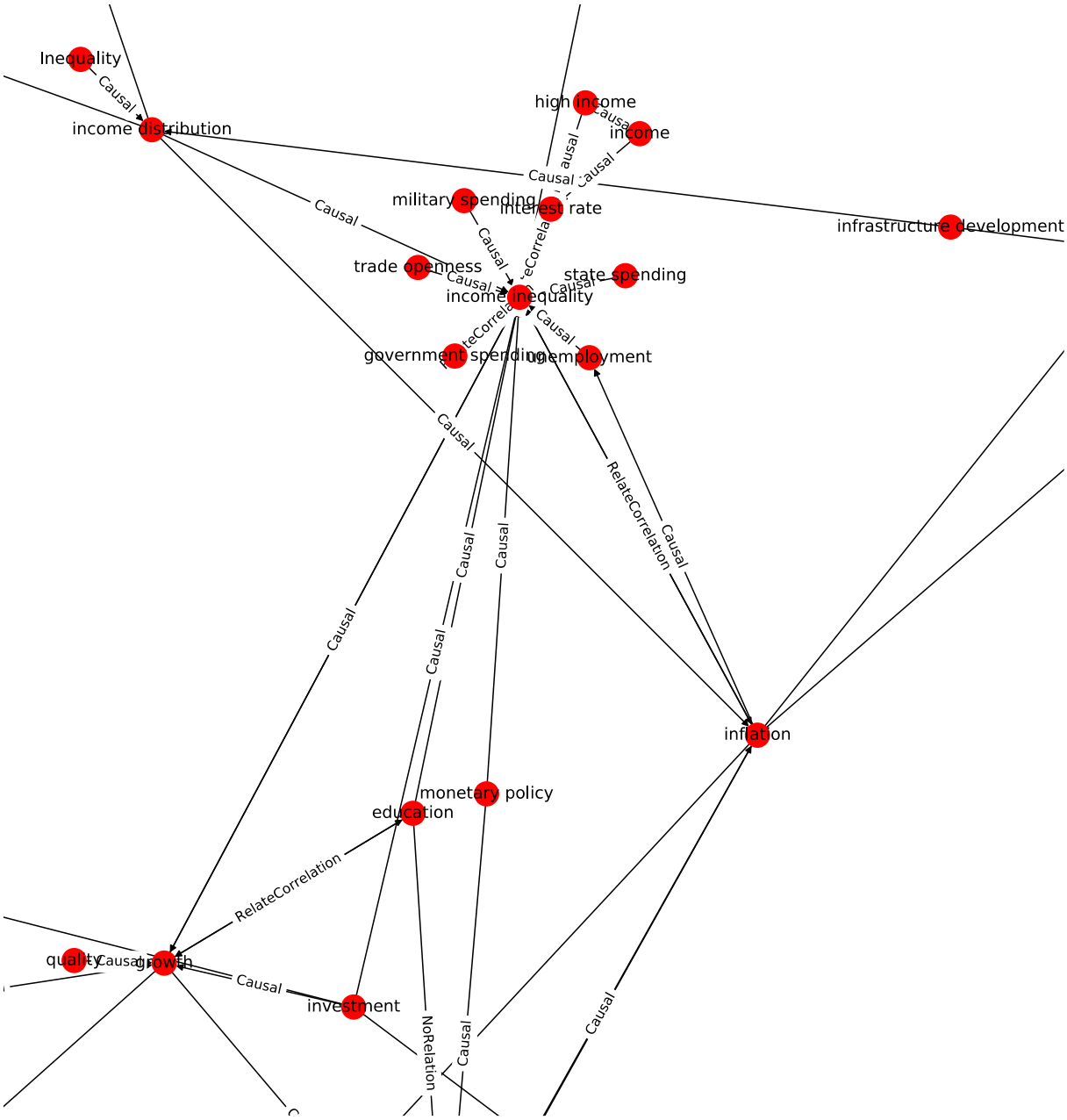


Figure 31: Zoom of figure 30

8. Conclusion

In this thesis, we have seen how to infer causal effects from causal diagrams. This theory has mainly been developed by J. Pearl and shows that we can deduce causal relationships from the topology of diagrams. The usefulness of those diagrams for economics has been debated and yet did not arrived at a consensus. The different ways of constructing a causal diagram have been examined. **This thesis focused on the construction of a causal diagram from scientific literature in economics, in an automated way.** In the academic world this has been studied mainly in the life-science literature and has been underexploited in economics.

The continued **growth of scientific papers** in economics results in researchers no longer able to 'process' all the information by reading it. With artificial intelligence, the creation of a graphical representation of all economic entities and their relations can be automated. Indeed, in recent years, natural language processing has seen tremendous breakthroughs. It is now possible to **transform human language in structured diagrams**. By doing so, researchers and analysts will have more time to study the resulting graph and decide which question promises to be relevant to address in their research and analysis.

Therefore, a causal diagram was created using IBM Watson tools. **Watson Knowledge Studio** was used to train a machine learning model to recognize economic variables and the relation between those variables in economic literature and more specifically in abstracts. Once the model created, it was deployed in the **Watson Natural Language Understanding (NLU)** service to create a diagram. 102 abstracts regarding the topic "income inequality" were labelled and used to train the model. Only **findings of papers** were extracted and 3 relations were considered: **causality, correlation and no relation**. Different versions of the model were created, they differ by number of annotated documents, the number of entries in the dictionary, the type system and the consistency of annotations. The best version is version 9 yields an **F1 score of 0.83 for entity recognition and 0.42 for relation recognition** in the test set. This is an increase of 0.32 for entity recognition and 0.15 for relation recognition in comparison with version 1. The main failures of the model reside in: failing to find relationships because it fails to identify correctly the entities in the first place, mixing causal relation and correlation, and relations that are annotated by the model and that do not represent findings of the paper. The low F1 score for relations extraction is not surprising considering the **small amount of training data** with in total around 13.000 words in the training set while the WKS documentation recommends having 300.000 words in the training set to yield sufficient results. Version 9 of the model was deployed to the NLU service and a diagram of the test set was created. Only the relations having a probability to be correct higher than 50% were represented.

This thesis shows that it is possible to extract causal relationships from abstracts in economics with a supervised machine learning model.

Several hypotheses can be made for causal diagrams to be underexploited in economics. Maybe **econometricians are scared that we will not use their tools** anymore. Causal diagrams should be complementary statistical tools, not aim to replace them. On the other hand, to my knowledge, **Pearl never discussed the construction of causal diagrams using NLP** which makes it possible to automatize the process. In comparison with the medical domain, it is easier to extract information from medical records because they often include **more structured text**. Moreover, it is **difficult for economists to use NLP because it often requires expertise** in informatics. Another explanation for the underuse of NLP to construct causal diagrams in economics could be that the **accuracy and precisions of relation extraction is still low**. It is for example difficult to extract conditional causality and errors can be present in the graph which makes them not reliable enough. Moreover, it is **time consuming for a researcher to create the model and train it.** Once, the precision of entities and relations extraction will be high enough, because it will rely on a model that is already trained for economics and that can just be used by researchers, maybe then we will be able to fully exploit this technology.

Future work could focus on: 1) Fixing some of the annotation's problems faced in this thesis such as, recognizing conditional independencies, find a way to annotated gathered variables or recognizing intra-sentential relationships. 2) Proof the usefulness of making causal diagrams in an automated way by comparing results using this method and results not using this method. 3) Improve the annotated dataset to make the model more efficient. This could be done by using active learning[7] to accelerate the construction of the training dataset. 4) Create a global causal diagram of economics that researchers could access and feed in an interactive way. 5) Include in the causal diagrams the methods used to find a specific causal relationship and link the relationships back to papers that examine them.

[7] Active learning = semi-supervised learning

9. References

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables: Rejoinder. Journal of the American Statistical Association, 91(434), 468.

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

Auer, S., & Mann, S. (2019). Towards an open research knowledge graph. *The Serials Librarian*, *76*(1-4), 35-41.

Auer, S., Kovtun, V., Prinz, M., Kasprzik, A., Stocker, M., & Vidal, M. E. (2018, June). Towards a knowledge graph for science. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics* (pp. 1-6).

Augenstein, I., Das, M., Riedel, S., Vikraman, L., & McCallum, A. (2017). Semeval 2017 task 10: Science ie-extracting keyphrases and relations from scientific publications. *arXiv preprint arXiv:1704.02853*.

Balashankar, A., Chakraborty, S., Fraiberger, S., & Subramanian, L. (2019, November). Identifying predictive causal factors from news streams. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 2338-2348).

Berisha E, Meszaros J, Olson E (2018) Income inequality, equities, household debt, and interest rates: evidence from a century of data. J Int Money Finance 80:1–14

Blake, C. (2010). Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of biomedical informatics*, *43*(2), 173-189.

Blyth, C. R. (1972). On Simpson's Paradox and the Sure-Thing Principle. Journal of the American Statistical Association, 67(338), 364–366. https://doi.org/10.1080/01621459.1972.10482387

Bucevska, V. (2019). Determinants of Income Inequality in EU Candidate Countries: A Panel Analysis. *Economic Themes*.

Bui, Q. C., Nualláin, B. Ó., Boucher, C. A., & Sloot, P. M. (2010). Extracting causal relations on HIV drug resistance from literature. *BMC bioinformatics*, *11*(1), 1-11.

Card, D., & DellaVigna, S. (2014). Page limits on economics articles: Evidence from two journals. *Journal of Economic Perspectives*, *28*(3), 149-68.

Checchi, D., & García-Peñalosa, C. (2010). Labour market institutions and the personal distribution of income in the OECD. *Economica*, *77*(307), 413-450.

Chen, H., & Luo, X. (2019). An automatic literature knowledge graph and reasoning network modeling framework based on ontology and natural language processing. *Advanced Engineering Informatics*, *42*, 100959.

Chen, V. Z., Montano-Campos, F., & Zadrozny, W. (2020). Causal Knowledge Extraction from Scholarly Papers in Social Sciences. *arXiv preprint arXiv:2006.08904*.

Chetverikov, D., Santos, A., & Shaikh, A. M. (2018). The Econometrics of Shape Restrictions. *Annual Review of Economics*, *10*(1), 31–63.

Chiticariu, L., Li, Y., & Reiss, F. (2013, October). Rule-based information extraction is dead! long live rule-based information extraction systems!. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 827-832).

Cox, D. R., & Wermuth, N. (1995). Causal diagrams for empirical research: Discussion of 'Causal diagrams for empirical research'by J. Pearl. *Biometrika*, *82*(4), 688-689.

d'Hombres, B., Weber, A., & Elia, L. (2012). Literature review on income inequality and the effects on social outcomes. *JRC Scientific and Policy Reports*.

de Haan, J., and Sturm, J. (2016), 'Finance and Income Inequality: A Review and New Evidence', CESifo Working Paper Series No. 6079.

DEFT (n.d). Semantic similarity and fine-grained information extraction. Retrieved 2021, November 12 from https://deft.limsi.fr/2020/index-en.html

Derchi, G., Visentin, M., Marchio, V., Lardani, L., Barone, A., Prenassi, M., & Marceglia, S. (2020). Application of IBM Watson to Support Literature Reviews: A Preliminary Experience in Restorative Dentistry. *Digital Personalized Health and Medicine*, 1201-1202.

Dessì, D., Osborne, F., Recupero, D. R., Buscaldi, D., & Motta, E. (2021). Generating knowledge graphs by employing Natural Language Processing and Machine Learning techniques within the scholarly domain. *Future Generation Computer Systems*, *116*, 253-264.

Evans, D. (2021, January). How to write the abstract of your development economics paper. Center from global development. Retrieved 2021, December 10 from: https://www.cgdev.org/blog/how-write-abstract-your-development-economics-paper#:~:text=The%20Journal%20of%20Development%20Economics,them%20for%206%2D7%20sentences.

Florian, R., Hassan, H., Ittycheriah, A., Jing, H., Kambhatla, N., Luo, X., ... & Roukos, S. (2004). *A statistical model for multilingual entity detection and tracking*. IBM THOMAS J WATSON RESEARCH CENTER YORKTOWN HEIGHTS NY.

Forrester (2021, February). The Total Economic Impact Of IBM Watson Natural Language Processing (NLP) Solutions. https://www.ibm.com/downloads/cas/XMRMP7XK

Freedman, D. (1995). Causal diagrams for empirical research: Discussion of 'causal diagrams for empirical research'by J. Pearl. *Biometrika*, *82*(4), 692-693.

Fritzner, J. E. H. (2017). *Automated information extraction in natural language* (Master's thesis, NTNU).

Furceri, D., & Ostry, J. D. (2019). Robust determinants of income inequality. *Oxford Review of Economic Policy*, *35*(3), 490-517.

Gábor, K., Buscaldi, D., Schumann, A. K., QasemiZadeh, B., Zargayouna, H., & Charnois, T. (2018, June). Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation* (pp. 679-688).

Geiger, D., Verma, T., & Pearl, J. (1990). Identifying independence in Bayesian networks. Networks, 20(5), 507–534.

Georgescu, T. M. (2020). Natural language processing model for automatic analysis of cybersecurity-related documents. *Symmetry*, *12*(3), 354.

Gerdtham, U.-G., & Johannesson, M. (2004). Absolute Income, Relative Income, Income Inequality, and Mortality. *Journal of Human Resources*, *39*(1), 228-247.

Github. (n.d.). IBM/build-knowledge-base-with-domain-specific-documents. Retrieved 2021 September 13. From: https://github.com/IBM/build-knowledge-base-with-domain-specific-documents

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in genetics*, *10*, 524.

Gu, J., Qian, L., & Zhou, G. (2016). Chemical-induced disease relation extraction with various linguistic features. *Database*, *2016*. doi:10.1093/database/baw042

Hailemariam, A., Sakutukwa, T., & Dzhumashev, R. (2021). Long-term determinants of income inequality: evidence from panel data over 1870–2016. *Empirical Economics*, *61*(4), 1935-1958.

Haughton, J., & Khandker, S. R. (2009). *Handbook on poverty+ inequality*. World Bank Publications.

Heinze-Deml, C., Maathuis, M. H., & Meinshausen, N. (2018). Causal structure learning. *Annual Review of Statistics and Its Application*, *5*, 371-391.

Huang, F., Ahuja, A., Downey, D., Yang, Y., Guo, Y., & Yates, A. (2014). Learning representations for weakly supervised natural language processing tasks. *Computational Linguistics*, *40*(1), 85-120.

Huang, Y., & Valtorta, M. (2012). Pearl's calculus of intervention is complete. *arXiv preprint arXiv:1206.6831*.

Hubbard, K. E., & Dunbar, S. D. (2017). Perceptions of scientific research literature and strategies for reading papers depend on academic career stage. *PloS one*, *12*(12), e0189753.

Hünermund, *P (2021). Causal Inference with Directed Acyclic Graphs [MOOC)]. From, https://p-hunermund.com/teaching/*

Ichim, A., Neculita, M., & Sarpe, D. A. (2018). Income inequality. Literature review. *Risk in Contemporary Economy*, 252-259.

IDC (2020, Aout). Forecasts Strong 12.3% Growth for AI Market in 2020 Amidst Challenging Circumstances. https://www.idc.com/getdoc.jsp?containerId=prUS46757920

Imbens, G (2014a, November) "I *am glad that in his discussion of my paper Judea Pearl provides a link to the paper. I suggest the reader look at the paper and decide for him or herself whether it reflects the attitudes Judea ascribes to me…."* [Comment on the post " *Are economists smarter than epidemiologists?*"]. Retrieved August 15, 2021, from http://causality.cs.ucla.edu/blog/index.php/2014/10/27/are-economists-smarter-than-epidemiologists-comments-on-imbenss-recent-paper/

Imbens, G (2014b) Rejoinder of "Instrumental variables: an Econometrician Perspective". arXiv:1410.0482

Imbens, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, *58*(4), 1129-79.

Imbens, G. W., & Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, *62*(2), 467. https://doi.org/10.2307/2951620

Imbens, G. W., & Rubin, D. B. (1995). Causal diagrams for empirical research: Discussion of 'Causal diagrams for empirical research'by J. Pearl. *Biometrika*, *82*(4), 694-695.

Izumi, K., & Sakaji, H. (2019, August). Economic causal-chain search using text mining technology. In *International Joint Conference on Artificial Intelligence* (pp. 23-35). Springer, Cham.

Javapoint (n.d). Supervised machine learning. Retrieved 2021, October 18 from https://www.javatpoint.com/supervised-machine-learning

Joachims, T. (2002). *Learning to classify text using support vector machines* (Vol. 668). Springer Science & Business Media.

Kejriwal, M. (2019). *Domain-specific knowledge graph construction*. Springer.

Khoo, C. S., Chan, S., & Niu, Y. (2000, October). Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of the 38th annual meeting of the association for computational linguistics* (pp. 336-343).

Khoo, C. S., Kornfilt, J., Oddy, R. N., & Myaeng, S. H. (1998). Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing*, *13*(4), 177-186.

Khoo, C., Chan, S., & Niu, Y. (2002). The many facets of the cause-effect relation. In *The Semantics of Relationships* (pp. 51-70). Springer, Dordrecht.

Kim, S., Choi, S., & Seok, J. (2021, August). Keyword Extraction in Economics Literatures using Natural Language Processing. In *2021 Twelfth International Conference on Ubiquitous and Future Networks (ICUFN)* (pp. 75-77). IEEE.

Kitano, H. (2016). Artificial intelligence to win the nobel prize and beyond: Creating the engine for scientific discovery. *AI magazine*, *37*(1), 39-49.

Klein, D. (2005). *The unsupervised learning of natural language structure*. Stanford University.

Laiq, M., & Dieste, O. (2020, August). Chatbot-based interview simulator: A feasible approach to train novice requirements engineers. In *2020 10th International Workshop on Requirements Engineering Education and Training (REET)* (pp. 1-8). IEEE.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., ... & Bizer, C. (2015). Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, *6*(2), 167-195.

Li, A., Wang, X., Wang, W., Zhang, A., & Li, B. (2019, August). A survey of relation extraction of knowledge graphs. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data* (pp. 52-66). Springer, Cham.

Loken, K. V. (2007, January). Family Income and Children's Education: Using the Norwegian Oil Boom as a Natural Experiment. *Labour Economics*, *17*(1), 118-129.

Maddala, G. S. (1986). *Limited-dependent and qualitative variables in econometrics* (No. 3). Cambridge university press.

Matzkin, R. L. (1991). Semiparametric Estimation of Monotone and Concave Utility Functions for Polychotomous Choice Models. *Econometrica*, *59*(5), 1315.

Mayo, M (2018 October). The main approaches to Natural Language Processing Tasks. Retrieved 2021, December 15 from https://www.kdnuggets.com/2018/10/main-approaches-natural-language-processing-tasks.html

Mdingi, K., & Ho, S. Y. (2021). Literature Review on Income Inequality and Economic Growth. *MethodsX*, 101402.

Mihăilă, C., & Ananiadou, S. (2014). Semi-supervised learning of causal relations in biomedical scientific discourse. *Biomedical engineering online*, *13*(2), 1-24.

Monnin, P. (2014). Inflation and income inequality in developed economies. *CEP Working Paper Series*.

NLU Documentation (n.d.). Getting started with Natural Language Understanding. Retrieved, 2021 September 5 from: https://cloud.ibm.com/docs/natural-language-understanding?topic=natural-language-understanding-getting-started

NLU Presentation (n.d.). Watson Natural Language Understanding. Retrieved, 2021 September 5 from: https://www.ibm.com/cloud/watson-natural-language-understanding

Nordon, G., Koren, G., Shalev, V., Kimelfeld, B., Shalit, U., & Radinsky, K. (2019, July). Building causal graphs from medical literature and electronic medical records. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 1102-1109).

Nuzzolese, A. G., Gentile, A. L., Presutti, V., & Gangemi, A. (2016, October). Conference linked data: the scholarlydata project. In *International Semantic Web Conference* (pp. 150-158). Springer, Cham.

Pearl, J (2014a, October). *Are economists smarter than epidemiologists? (Comments on Imbens's recent paper).* Retrieved August 15, 2021, from http://causality.cs.ucla.edu/blog/index.php/2014/10/27/are-economists-smarter-than-epidemiologists-comments-on-imbenss-recent-paper/

Pearl, J (2020, January). *On Imbens's Comparison of Two Approaches to Empirical Economics*. Retrieved September 17, 2021, from http://causality.cs.ucla.edu/blog/index.php/2020/01/29/on-imbens-comparison-of-two-approaches-to-empirical-economics/

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669-688.

Pearl, J. (2000). Causality: models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, *19*.

Pearl, J., & Bareinboim, E. (2011). *Transportability across studies: A formal approach*. CALIFORNIA UNIV LOS ANGELES DEPT OF COMPUTER SCIENCE.

Pearl, J., & Bareinboim, E. (2014b). External validity: From do-calculus to transportability across populations. *Statistical Science*, *29*(4), 579-595.

Pearl, J., & Mackenzie, D. (2018). The book of why: the new science of cause and effect. Basic books.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Persson, T., & Tabellini, G. (1994). ls lnequality Harmful for Growth. *American Economic Review*, *84*(3), 600-621.

Royan, C., Langé, J. M., & Abidi, Z. (2020). Extraction d'information de spécialité avec un système commercial générique. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes* (pp. 79-90). ATALA; AFCP.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., & Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, *308*(5721), 523-529.

Sekhon, J. S., & Shem-Tov, Y. (2020). Inference on a New Class of Sample Average Treatment Effects. *Journal of the American Statistical Association*, *116*(534), 798–804.

Shpitser, I., & Pearl, J. (2006, February). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence* (Vol. 21, No. 2, p. 1219). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Singh, G., Papoutsoglou, E. A., Keijts-Lalleman, F., Vencheva, B., Rice, M., Visser, R. G., ... & Finkers, R. (2021). Extracting knowledge networks from plant scientific literature: potato tuber flesh color as an exemplary trait. *BMC plant biology*, *21*(1), 1-14.

Song, F., Parekh, S., Hooper, L., Loke, Y. K., Ryder, J., Sutton, A. J., ... & Harvey, I. (2010). Dissemination and publication of research findings: an updated review of related biases. *Health technology assessment*, *14*(8), 1-220.

Sorgente, A., Vettigli, G., & Mele, F. (2013). Automatic extraction of cause-effect relations in Natural Language Text. *DART@ AI* IA*, *2013*, 37-48.

Spirtes, P., Glymour, C. N., Scheines, R., & Heckerman, D. (2000). *Causation, prediction, and search*. MIT press.

Stock, J. H., & Watson, M. W. (2015). Introduction to econometrics 3rd ed.

Tenopir, C., King, D. W., Edwards, S., & Wu, L. (2009, January). Electronic journals and changes in scholarly article seeking and reading patterns. In *Aslib proceedings*. Emerald Group Publishing Limited.

Tilly, S., & Livan, G. (2021). Macroeconomic forecasting with statistically validated knowledge graphs. *arXiv preprint arXiv:2104.10457*.

Tonin, L. (2017). Annotating mentions of coronary artery disease in medical reports (Master's thesis, KTH)

Valenzuela-Escárcega, M. A., Babur, Ö., Hahn-Powell, G., Bell, D., Hicks, T., Noriega-Atala, E., ... & Morrison, C. T. (2018). Large-scale automated machine reading discovers new cancer-driving mechanisms. *Database*, *2018*.

Wang, K., Shen, Z., Huang, C., Wu, C. H., Dong, Y., & Kanakia, A. (2020). Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, *1*(1), 396-413.

White, K. E., Robbins, C., Khan, B., & Freyman, C. (2017). Science and engineering publication output trends: 2014 shows rise of developing country output while developed countries dominate highly cited publications. *National Center for Science and Engineering Statistics InfoBrief*, 1-7.

WKS documentation. (n.d). Getting started with Knowledge Studio. Retrieved, 2021 September 14 from: https://cloud.ibm.com/docs/watson-knowledge-studio?topic=watson-knowledge-studio-wks_tutintro

WKS Presentation (n.d.). IBM Watson Knowledge Studio. Retrieved, 2021 September 10 from: https://www.ibm.com/cloud/watson-knowledge-studio

Yang, J., Han, S. C., & Poon, J. (2021). A survey on extraction of causal relations from natural language text. *arXiv preprint arXiv:2101.06426*.

Yang, Y., Pang, Y., & Huang, G. (2020). The Knowledge Graph for Macroeconomic Analysis with Alternative Big Data. *arXiv preprint arXiv:2010.05172*.

Yu, B., Li, Y., & Wang, J. (2019, November). Detecting causal language use in science findings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 4664-4674)

10. Appendix

10.1. Annotation guidelines

Types of entities: variables, evidence, relation, negation

Types of relation: causation, correlation, no relation

- Relation words

Causation indicator	Correlation indicator	No relation indicator
Account for	Associated with/between/	Not explain
Affect	Among	No effect
Associated with	Connection between	Limited effect
Boosted by	Correlated	
Causal relationship	Links between	
Caused by	Matter for	
Conditioned by	Relate	
Contribute to	Relationship between	
Depends on	Rest upon	
Depress		
Driven by changes in		
Effect on/of		
Effective in		
Engine of		
Exacerbated		
Explains/ explained by		
Impact on/of		
Influence		

Influence on/of Key driver Lead to One-way relationship Plays a roll in Reduce, rising, increase, Decreasing, higher Result in/resulting from Reviews Sensitive to The greater …. The lower…		

- Evidence words

Don't take relation into account	Take the relation into account
Analyze/ analysis	Argue
Assess	Conclude/ conclusion
Confirms	Demonstrate
Develop	Develop
Discuss	Discussion
Estimate	Document
Evaluation	Estimates
Examine	Evidence
Explain	Evidence-based
Explore	Findings/find/…
Hypothesis	Identify
Investigate	Predict
Look at	Presented
Method	Result
No evidence	Show/shown
Note	Suggest
Objective	
Present	
Purpose	
Seeks evidence	
Starting	
Study	
Test	
Uses	

10.2. Results of the different versions

Version 1:

	Entity Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
	Variable	0.55	0.55	0.55	64% (80/125)	12% (80/664)	100% (5/5)
	evidence	0.67	0.8	0.57	8% (10/125)	2% (10/664)	40% (2/5)
⚠	**relation**	**0.4**	**0.45**	**0.36**	**28% (35/125)**	**5% (35/664)**	**100% (5/5)**
	Overall Statistics	0.51	0.54	0.49	100% (125/125)	19% (125/664)	100% (5/5)

	Relation Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
⚠	**Causal**	**0.14**	**0.13**	**0.17**	**56% (25/44)**	**4% (25/664)**	**40% (2/5)**
⚠	**NoRelation**	**0**	**0**	**0**	**7% (3/44)**	**0% (3/664)**	**20% (1/5)**
	RelateCorrelation	0.57	1	0.4	36% (16/44)	2% (16/664)	60% (3/5)
	Overall Statistics	0.27	0.3	0.25	100% (44/44)	7% (44/664)	60% (3/5)

Version 2:

	Entity Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
	Variable	0.69	0.76	0.63	56% (54/94)	16% (54/343)	100% (3/3)
	evidence	0.67	0.8	0.57	11% (10/94)	3% (10/343)	67% (2/3)
	relation	0.63	0.73	0.55	32% (30/94)	9% (30/343)	100% (3/3)
	Overall Statistics	0.67	0.76	0.6	100% (94/94)	27% (94/343)	100% (3/3)

	Relation Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
⚠	**Causal**	**0.25**	**0.2**	**0.33**	**56% (25/44)**	**7% (25/343)**	**67% (2/3)**
⚠	**NoRelation**	**0**	**0**	**0**	**7% (3/44)**	**1% (3/343)**	**33% (1/3)**
	RelateCorrelation	0.57	1	0.4	36% (16/44)	5% (16/343)	100% (3/3)
	Overall Statistics	0.33	0.33	0.33	100% (44/44)	13% (44/343)	100% (3/3)

Version 3:

	Entity Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
	Variable	0.71	0.77	0.67	55% (54/98)	16% (54/343)	100% (3/3)
	evidence	0.63	0.83	0.5	13% (13/98)	4% (13/343)	100% (3/3)
	relation	0.7	0.76	0.65	32% (31/98)	9% (31/343)	100% (3/3)
	Overall Statistics	0.7	0.78	0.63	100% (98/98)	28% (98/343)	100% (3/3)

	Relation Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
⚠	**Causal**	**0.31**	**0.29**	**0.33**	**56% (25/44)**	**7% (25/343)**	**67% (2/3)**
⚠	**NoRelation**	**0**	**0**	**0**	**7% (3/44)**	**1% (3/343)**	**33% (1/3)**
	RelateCorrelation	0.57	1	0.4	36% (16/44)	5% (16/343)	100% (3/3)
	Overall Statistics	0.38	0.44	0.33	100% (44/44)	13% (44/343)	100% (3/3)

Version 4:

	Entity Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
	Variable	0.75	0.77	0.73	65% (193/295)	17% (193/1157)	100% (10/10)
	evidence	0.73	1	0.57	9% (28/295)	2% (28/1157)	90% (9/10)
	relation	0.72	0.77	0.67	25% (74/295)	6% (74/1157)	100% (10/10)
	Overall Statistics	0.74	0.79	0.7	100% (295/295)	25% (295/1157)	100% (10/10)

	Relation Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
⚠	**Causal**	**0.15**	**0.12**	**0.19**	**40% (47/118)**	**4% (47/1157)**	**60% (6/10)**
⚠	**NoRelation**	**0**	**0**	**0**	**26% (31/118)**	**3% (31/1157)**	**30% (3/10)**
⚠	**RelateCorrelation**	**0.19**	**0.25**	**0.15**	**34% (40/118)**	**3% (40/1157)**	**50% (5/10)**
	Overall Statistics	0.14	0.15	0.14	100% (118/118)	10% (118/1157)	100% (10/10)

Version 5:

	Entity Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
	Variable	0.8	0.81	0.79	66% (100/151)	17% (100/601)	100% (5/5)
	evidence	0.82	1	0.69	9% (13/151)	2% (13/601)	80% (4/5)
	relation	0.79	0.83	0.76	25% (38/151)	6% (38/601)	100% (5/5)
	Overall Statistics	0.8	0.84	0.77	100% (151/151)	25% (151/601)	100% (5/5)

	Relation Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
⚠	**Causal**	**0.45**	**0.33**	**0.71**	**32% (20/63)**	**3% (20/601)**	**60% (3/5)**
⚠	**NoRelation**	**0**	**0**	**0**	**49% (31/63)**	**5% (31/601)**	**60% (3/5)**
⚠	**RelateCorrelation**	**0**	**0**	**0**	**19% (12/63)**	**2% (12/601)**	**40% (2/5)**
	Overall Statistics	0.28	0.29	0.26	100% (63/63)	10% (63/601)	100% (5/5)

Version 6:

Entity Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
Variable	0.79	0.82	0.76	67% (292/439)	16% (292/1803)	100% (14/14)
evidence	0.79	0.89	0.72	10% (43/439)	2% (43/1803)	93% (13/14)
relation	0.78	0.87	0.71	24% (104/439)	6% (104/1803)	100% (14/14)
Overall Statistics	0.79	0.84	0.74	100% (439/439)	24% (439/1803)	100% (14/14)

Relation Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
⚠ **Causal**	**0.39**	**0.34**	**0.45**	**61% (104/170)**	**6% (104/1803)**	**71% (10/14)**
⚠ **NoRelation**	**0**	**0**	**0**	**18% (31/170)**	**2% (31/1803)**	**21% (3/14)**
RelateCorrelation	0.63	0.75	0.55	21% (35/170)	2% (35/1803)	43% (6/14)
Overall Statistics	0.4	0.4	0.4	100% (170/170)	9% (170/1803)	100% (14/14)

Version 7:

Entity Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
Variable	0.79	0.82	0.76	65% (292/447)	16% (292/1803)	100% (14/14)
evidence	0.78	0.88	0.7	10% (43/447)	2% (43/1803)	93% (13/14)
⚠ **negation**	**0**	**0**	**0**	**2% (8/447)**	**0% (8/1803)**	**21% (3/14)**
relation	0.79	0.88	0.72	23% (104/447)	6% (104/1803)	100% (14/14)
Overall Statistics	0.77	0.83	0.72	100% (447/447)	25% (447/1803)	100% (14/14)

Relation Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
⚠ **Causal**	**0.39**	**0.35**	**0.45**	**61% (104/170)**	**6% (104/1803)**	**71% (10/14)**
⚠ **NoRelation**	**0.2**	**0.5**	**0.13**	**18% (31/170)**	**2% (31/1803)**	**21% (3/14)**
RelateCorrelation	0.63	0.75	0.55	21% (35/170)	2% (35/1803)	43% (6/14)
Overall Statistics	0.42	0.43	0.42	100% (170/170)	9% (170/1803)	100% (14/14)

Version 8

Entity Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
Variable	0.77	0.82	0.73	64% (570/893)	18% (570/3153)	100% (23/23)
evidence	0.8	0.93	0.7	11% (98/893)	3% (98/3153)	100% (23/23)
negation	**0.27**	**0.33**	**0.22**	**1% (12/893)**	**0% (12/3153)**	**26% (6/23)**
relation	0.87	0.95	0.8	24% (213/893)	7% (213/3153)	100% (23/23)
Overall Statistics	0.79	0.86	0.74	100% (893/893)	28% (893/3153)	100% (23/23)

Relation Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
Causal	**0.36**	**0.39**	**0.34**	**66% (215/325)**	**7% (215/3153)**	**74% (17/23)**
NoRelation	**0**	**0**	**0**	**11% (35/325)**	**1% (35/3153)**	**22% (5/23)**
RelateCorrelat...	**0.4**	**0.4**	**0.4**	**23% (75/325)**	**2% (75/3153)**	**48% (11/23)**
Overall Statistics	0.34	0.36	0.31	100% (325/325)	10% (325/3153)	100% (23/23)

Version 9

Entity Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
Variable	0.81	0.86	0.76	65% (593/914)	19% (593/3153)	100% (23/23)
evidence	0.84	0.94	0.77	11% (99/914)	3% (99/3153)	100% (23/23)
negation	0.57	0.8	0.44	1% (9/914)	0% (9/3153)	26% (6/23)
relation	0.86	0.95	0.79	23% (213/914)	7% (213/3153)	100% (23/23)
Overall Statistics	0.83	0.89	0.77	100% (914/914)	28% (914/3153)	100% (23/23)

Relation Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
Causal	**0.38**	**0.41**	**0.36**	**66% (225/339)**	**7% (225/3153)**	**78% (18/23)**
NoRelation	**0**	**0**	**0**	**10% (35/339)**	**1% (35/3153)**	**22% (5/23)**
RelateCorrelat...	**0.44**	**0.45**	**0.43**	**23% (79/339)**	**3% (79/3153)**	**48% (11/23)**
Overall Statistics	0.36	0.39	0.33	100% (339/339)	11% (339/3153)	100% (23/23)

10.3. Graphs of results

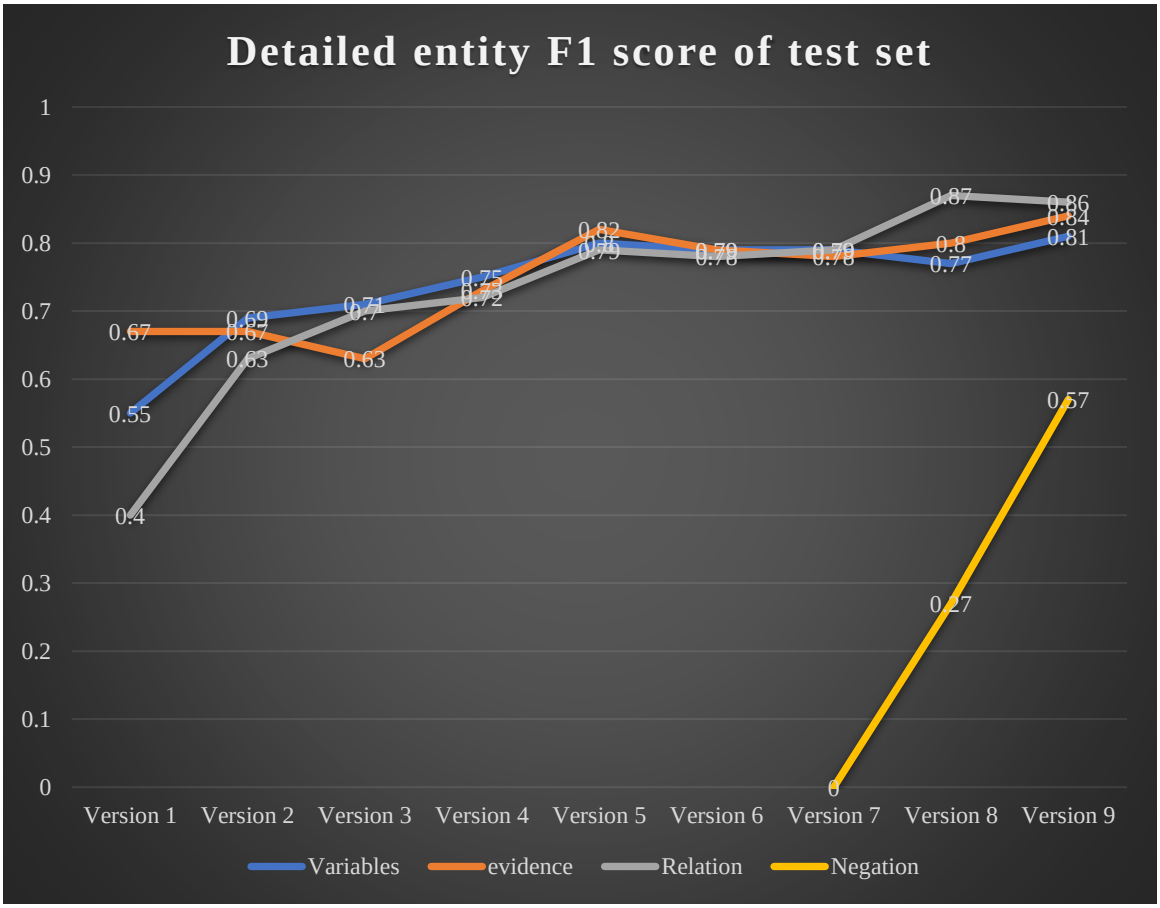


Figure 32: Detailed entity F1 score of test set

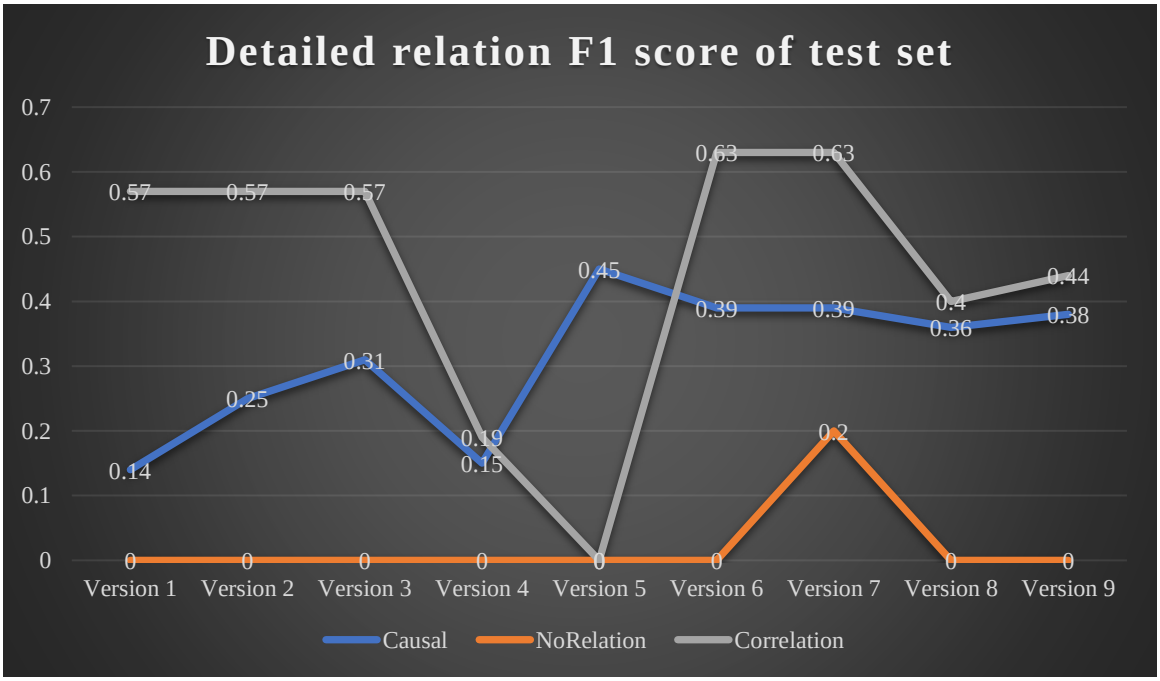


Figure 33: Detailed relation F1 score of test set

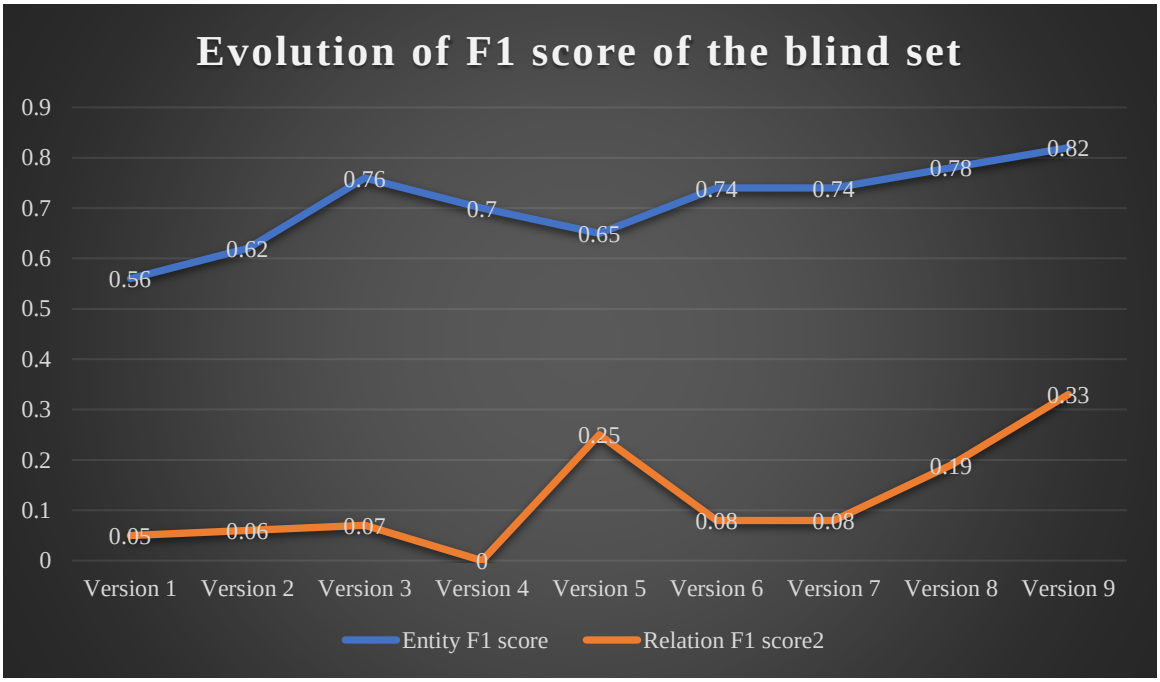


Figure 34: Evolution of the F1 score of the blind set

10.4. Causal diagrams

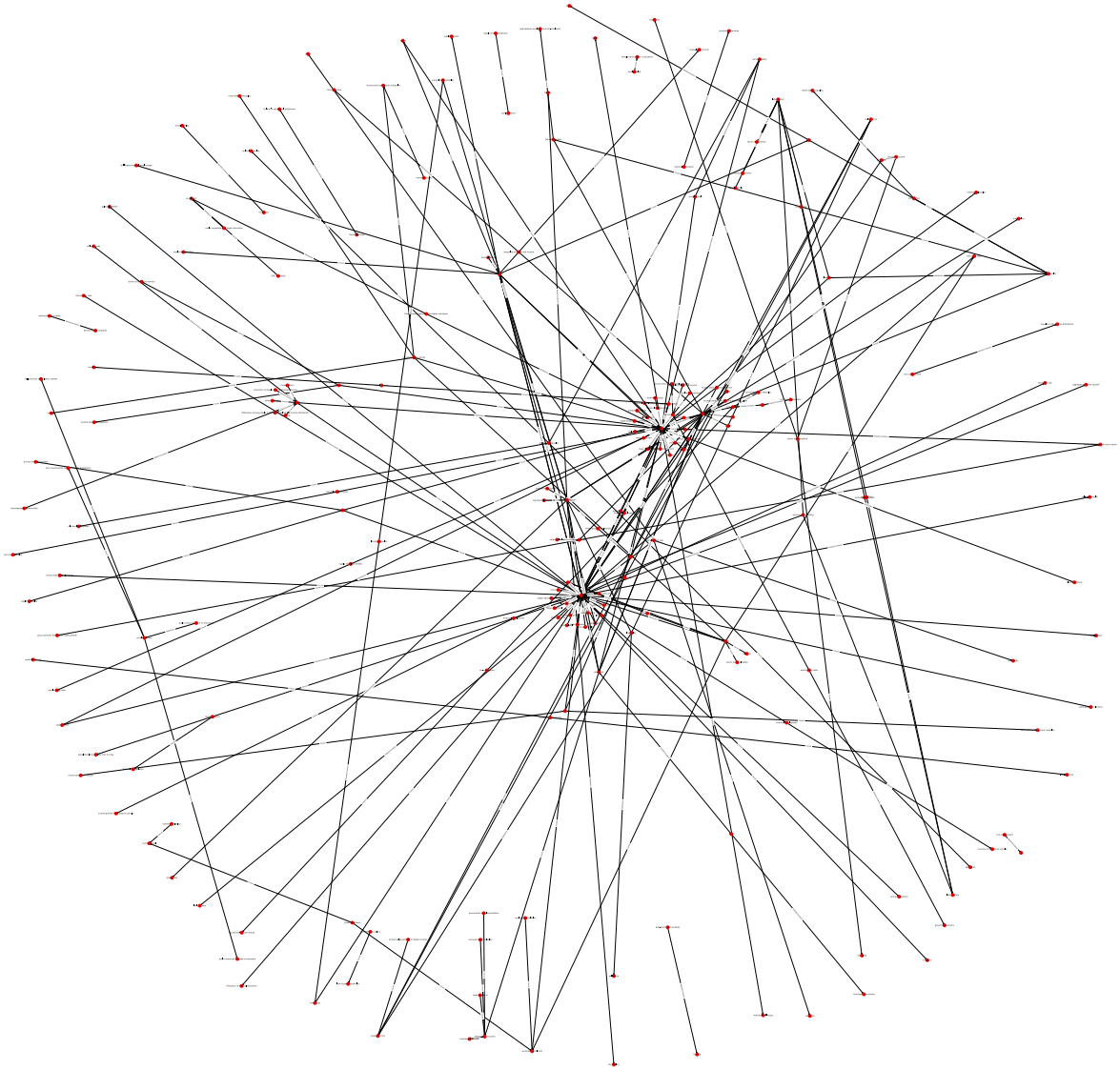


Figure 35: Causal diagram extracted from all the abstracts

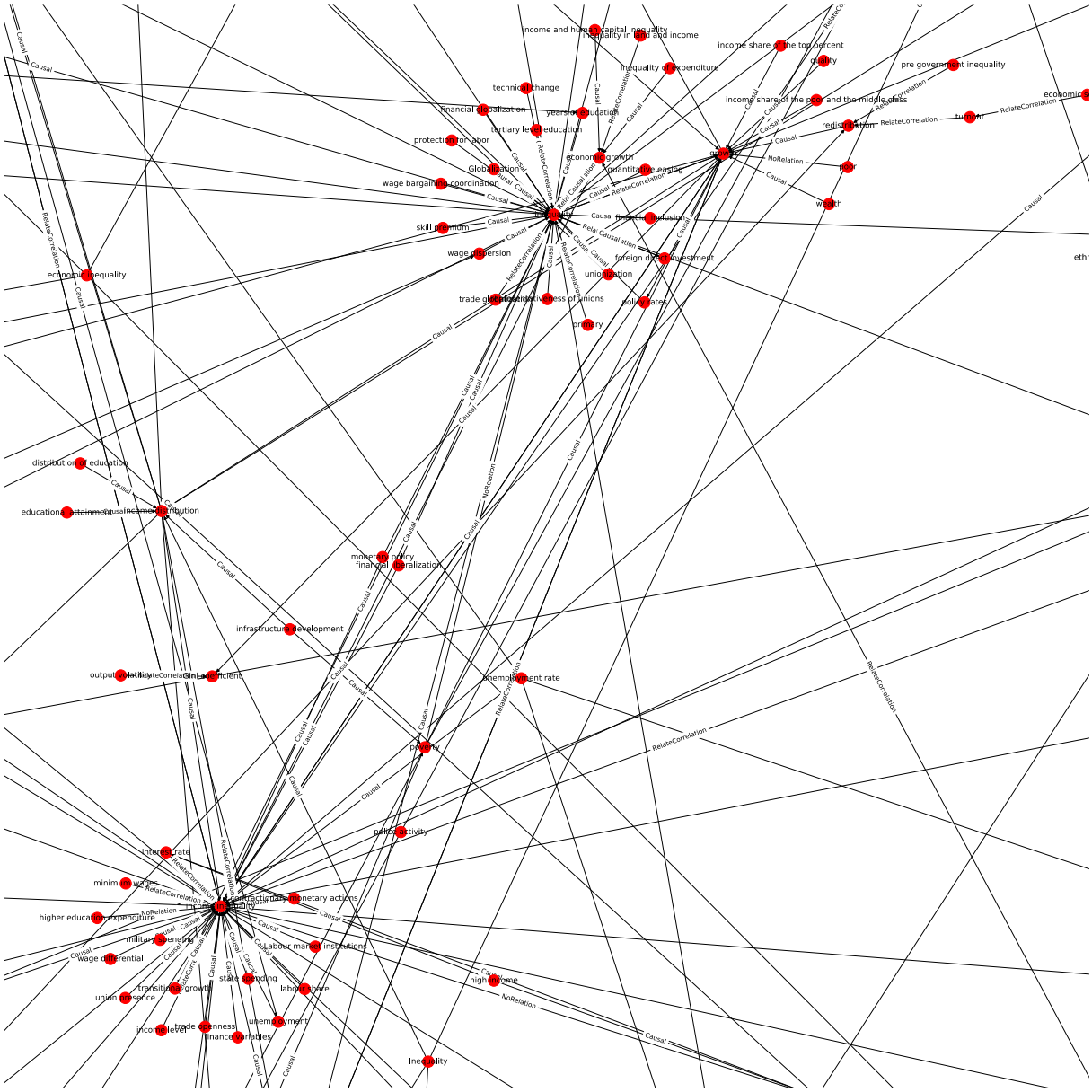


Figure 36: Zoom of figure 35

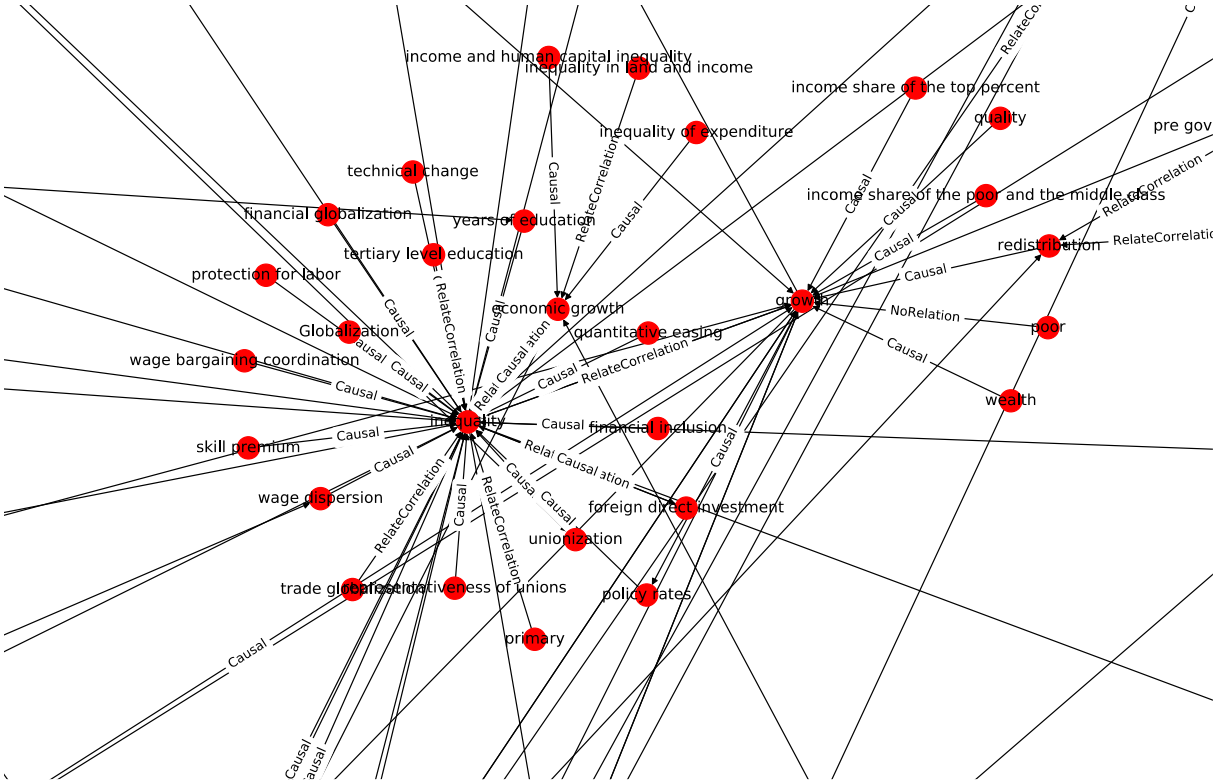


Figure 37: Zoom of figure 36