*Natural Language Engineering* (2022), **28**, pp. 361–400
doi:10.1017/S135132492100036X

CAMBRIDGE
UNIVERSITY PRESS

SURVEY PAPER

# A survey of the extraction and applications of causal relations

Brett Drury[1,*], Hugo Gonçalo Oliveira[2] and Alneu de Andrade Lopes[3]

[1]LIAAD INESC Tec, R. Dr. Roberto Frias, Porto, Portugal, [2]CISUC, Department of Informatics Engineering, Universidade de Coimbra, Coimbra, Portugal and [3]ICMC, USP Av. Trab. São Carlense, São Paulo, Brazil
*Corresponding author. E-mail: brett.drury@gmail.com

(Received 28 December 2019; revised 31 October 2021; accepted 1 November 2021; first published online 20 January 2021)

**Abstract**

Causation in written natural language can express a strong relationship between events and facts. Causation in the written form can be referred to as a causal relation where a cause event entails the occurrence of an effect event. A cause and effect relationship is stronger than a correlation between events, and therefore aggregated causal relations extracted from large corpora can be used in numerous applications such as question-answering and summarisation to produce superior results than traditional approaches. Techniques like logical consequence allow causal relations to be used in niche practical applications such as event prediction which is useful for diverse domains such as security and finance. Until recently, the use of causal relations was a relatively unpopular technique because the causal relation extraction techniques were problematic, and the relations returned were incomplete, error prone or simplistic. The recent adoption of language models and improved relation extractors for natural language such as Transformer-XL (Dai *et al.* (2019). *Transformer-xl: Attentive language models beyond a fixed-length context*. arXiv preprint arXiv:1901.02860) has seen a surge of research interest in the possibilities of using causal relations in practical applications. Until now, there has not been an extensive survey of the practical applications of causal relations; therefore, this survey is intended precisely to demonstrate the potential of causal relations. It is a comprehensive survey of the work on the extraction of causal relations and their applications, while also discussing the nature of causation and its representation in text.

**Keywords:** Causal relations; Survey; Sentiment; Event prediction; Information retrieval; Cause identification

## 1. Introduction

Causation is a temporal relationship where a cause event forces the occurrence of an effect event at a later point in time. An event can be seen as things ‘that develop and change fast in time’ (Casati and Varzi 2020). This definition can be applied to several things, but for this article, it can be assumed that an event can be an event or state. An event is an action where an actor intervenes in a previous stable system, such as a throwing a stone at a window. The window is in a stable state, and the action of throwing a stone at it disrupts its stable state by breaking it. A state is a condition of something that causes something to occur, such as a broken window causing a drop in temperature in a house. However, the state is likely to have been caused by a previous action such as the aforementioned breaking of a window.

Because text is used to record knowledge and human expression, causal relationships are often unwittingly represented in natural language. However, since natural language can be imprecise and lack the rigour of mathematical notation, causation in text can be expressed in several

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

different ways (Degand 1994). This ambiguity may increase the difficulty of extracting causation automatically from textual sources.

The inexactness of causation expressed in text has not inhibited efforts to detect causation in texts automatically using information extraction techniques. Causation in text is often referred to as causal relations and can have applications in areas such as predicting future events (Radinsky and Horvitz 2013), identifying causes of marine accidents (Tirunagari *et al.* 2012), question-answering (Girju 2003) and improving discourse parsers by adding CONTRAST and EXPLANATION–EVIDENCE relations (Marcu and Echihabi 2002).

Causation extraction from text and its application is not currently a popular research topic, and therefore this survey is intended to provide an introduction to the area as well as provide a state-of-the-art review. The scope of this article is to demonstrate the potential of causal relations rather than the nature of causation. However, the survey will discuss the nature of causality, but it should not be taken as a thorough and detailed discussion of the area.

The selection policy for papers was to select peer-reviewed papers and occasionally papers from preprint servers and high-quality technical reports. The initial search used major academic indexes such as Google Scholar[a] and Citeseer.[b] Terms such as 'causal relations' were used in the initial search, and papers published by companies on Beal's list[c] were removed. This set of articles were expanded by finding papers that cited articles in the original article set. This set of articles were filtered to remove low quality and non-peer-reviewed articles. The literature review was conducted between May 2018 and June 2019. Still, some papers from 2020 and 2021 were included, following referees' suggestions during the review process.

In Section 2, the survey provides a background to causation and its representation in natural language. Section 3 discusses causal information extraction, and in Section 3.8 a closely related area, emotion cause detection is discussed. In Section 4, the applications of causation are analysed. Finally, a conclusion is drawn on the future direction of research.

## 2. Defining causation

The initial work on causation was conducted by philosophers, who attempted to define the nature of causation. The consensus of the philosophical approach to causation is that a causal relationship is between 'events, facts and objects' (Vendler 1967) and that the cause event, fact or object (cause) must occur before the effect event, fact or object (effect) (de Spinoza 1996; Khoo, Chan, and Niu 2002). The relationship between cause and effect is complex, and an in-depth discussion is beyond the scope of this article. For an in-depth discussion of causality, the interested reader can consult (Copley and Wolf 2015; Beebee, Hitchcock, and Menzies 2015). The relationship, however, can be simplified to the occurrence of a cause that will inevitably produce an effect (Mackie 1965; de Spinoza 1996; Mellor 1998). However, as with very simple generalisations, there are many cases that escape the simple definition. For example, the oft cited causal relation between smoking and cancer, will superficially fail when rare individuals who have a combination of luck and genetics that allow them to live to beyond their expected lifespan without acquiring cancer. These outliers dominated the smoking causes cancer public debate for decades. There are physical laws that demonstrate a more stable causal relationship such as Boyle's law; however, these relationships are only causal while the fundamental laws as humans' understand them are true which was not true in the early universe (Barrow and Silk 1980) and are likely to be different in hypothetical alternate universes (Adams 2008). This discussion is out of scope for this article; however, it does show the complexity of the subject, and for the majority of natural language processing (NLP) applications the very simple definition of 'the occurrence of a cause that will inevitably produce

[a] https://scholar.google.com.br/.

[b] http://citeseerx.ist.psu.edu.

[c] https://beallslist.net/.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

an effect' is sufficient because the subtlety of causation will be beyond the comprehension of any automated computer system at the current time.

Causes and effects may be general or specific (Mackie 1974). General causation can be seen as the establishment of a causative link between two classes of events over time, for example, 'Smoking causes lung cancer' (Hitchcock 1995). Specific causation can be seen as the establishment of a causative link between two individual events, that is, specific causation is an instantiation of general causation. An example of an instantiation of a general causative statement 'smoking causes cancer' could be 'David's Smoking caused him to develop lung cancer' (Hitchcock 1995).

The nature of causation has been extended to include the idea of transitivity where, through a chain of events where one event is causally dependent upon another, the original event at the start of the chain is the cause of the final event (Hall and Paul 2013). In this case, the intervening events act as an agent where the causal effect of the initial event is passed to the next event. An in-depth discussion of transitivity is beyond the scope of this survey; however, interested readers can consult Hall and Paul (2013).

The study of causation has also included the notion of causation properties of events that have not happened. This type of causation tends to be conditional, for example, 'if he had not smoked he would not have cancer'. In this statement, it could be argued that smoking was the cause of his hypothetical cancer. This type of causation is known as a counterfactual. An in-depth discussion of counterfactuals is beyond the scope of this paper; however, interested readers can consult Collins, Hall, and Paul (2004).

The discovery of causal connections between events has several proposed methodologies. In this brief introduction to causation, we will discuss two proposed high-level ideas for causal discovery, namely interventions and manipulations. The fundamental idea behind interventions is that causality between an intervention and an outcome can be inferred. For example, through the administration of a drug and a patient recovering from a disease that the drug is designed to cure. The randomised double-blind medical trial is probably the best-known form of intervention to discover causal relationships between intervention and outcome (Pearl and Mackenzie 2018). The difference between traditional probability notation and intervention is that traditional probability notation relies upon observation. To distinguish between observation and intervention, a new form of notation known as Do Notation (Pearl 2012) has been developed. It should be noted that the notion of Do Notation has been seen as controversial in the statistics community (Pearl and Mackenzie 2018).

Manipulation at its most basic assumes that the value of the effect variable will change if the value of the cause variable is changed (Psillos 2007). This theory, established by Woodward (2005), is linked to counterfactuals and allows the posing of what-if questions about causal relations. Although other philosophers have commented about manipulation, perhaps the clearest example of manipulation is given by Psillos (2007) who gave the example of the causal relationship of the volume and pressure of a gas. The what-if questions that could be answered are what would be the pressure of the gas at given volumes?

This introduction to causation is an incomplete discussion of the nature of causation, and there are many publications that the interested reader can consult, such as Hall and Paul (2013) and Collins *et al.* (2004) for a deeper discussion of the area.

Causality in written language follows the nature of causality as deduced by philosophers and is 'expressed directly through lexical causatives' (Copley and Martine 2015). The lexical causatives such as the verb 'cause' connect events where 'one event is the cause of the other' (Altenberg 1984). This type of relation is known as a 'causative relation' (Vendler 1967; Altenberg 1984) or 'causal relation' and is a popular term in the research literature. A causative relation has the following features: (i) encapsulate the dual members of the relationship; (ii) express the type of relationship of the relation's members and (iii) identify the members in a coherent sequence (Altenberg 1984).

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

Causal verbs are lexical causatives that form part of a causative relation. Causal verbs can be grouped into three main categories: 'simple, resultative and instrumental' (Girju 2003). A simple causative verb represents the causal bridge between the cause and effect events. For example, the verb 'cause' plays the role of a simple causal verb in the causative relation: 'smoking causes cancer'. A resultative causative verb provides a causal linkage as well as some or all of the description of the effect. The verb 'crowed' can be categorised as a resultative causal verb, as demonstrated by the phrase 'the roosters that scratch in the yard of Brastagi's best hotel crowed me awake that dawn a few months ago' (Levin 1986). An instrumental verb contains the cause as well as the causal linkage as demonstrated by the phrase: 'The criminal punched the victim to death', where the verb 'punched' behaves as an instrumental verb. There have been further groupings of causal verbs based upon their semantic properties (Levin 1993). It should be noted that causal properties of verbs may depend upon their semantic role, where one use of the verb is causal and in another it is non-causal.

Even though they are a popular research area for lexical causatives, causal verbs are not the only form of lexical causative. In English, a causative relationship in text can also be represented by: (i) conjunctions (e.g., *because*) (Lorenz 1999) and (ii) prepositional phrases (e.g., *due to*) (Degand 2000). These forms of causal representation are infrequent in the research literature when compared with causal verbs, but they can form the causal link in a causative relation.

Although Copley and Martine (2015) claim that lexical causatives form an integral part of a causal relation, there are claims that causal relations do not require them. They can be implied from the context and the sequence of events in a sentence. For example, 'It was a sweltering day, and I was sun burnt'. This fragment implies that the writer's sunburn was acquired from the sun.

Causation in written natural language is complex, and other forms of causal relations such as counterfactuals can be represented in text. However, this is out of scope for this survey. For an in-depth discussion of causation in language, the interested reader can consult Copley and Martine (2015), Neeleman *et al.* (2012).

## 3. Information extraction

Applications and techniques that rely on causal information extracted from text depend upon an information extraction step that acquires causal relations from a collection of documents. For the purposes of this paper, this step will be referred to as causal relation extraction (CRE). The CRE literature search conducted for this paper revealed two major approaches: linguistic and machine learning.

It should be noted that a comprehensive CRE technique that can find the majority of the types of causation discovered and described by philosophers was not discovered in the research literature. The majority of the approaches tried to solve a simplified version of causation, which is cause event, causal verb and effect event. The subtle nature of causality has escaped automated methods of extraction.

The linguistic approach relies mainly upon rules or patterns which are constructed by humans. This section describes the major approaches to rule construction. The machine learning section describes data-driven approaches. Both approaches require data, therefore on Page 24 is a discussion of the various corpora available for CRE.

### *3.1 Linguistic*

The literature review found that the linguistic approach was the most frequent approach for CRE (Girju 2003; Chan and Lam 2005; Cole *et al.* 2006; Sakaji, Sekine, and Masuyama 2008b; Bui *et al.* 2010 Ishii, Ma, and Yoshikawa 2010a; Mulkar-Mehta *et al.* 2011a; Sadek 2013; Cao *et al.* 2014;

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

**Table 1.** Summary of linguistic approaches

| Approach | Description | Citations |
|---|---|---|
| Clue phrases | This approach uses a specific set of keywords (clues) to use in a rule-based approach. The clues are typically causal verbs, prepositions or continuations | Sakaji *et al.* (2008b), Cao *et al.* (2014), Ishii *et al.* (2010a), Mirza (2016) |
| Patterns | This approach uses frequently occurring patterns to extract causal relations | Girju (2003), Sadek (2013), Ishii *et al.* (2010a), Cole *et al.* (2006) |
| Parse tree | This approach uses parse trees and rules to identify causal relationships | Bui *et al.* (2010) |
| Logical rules | This approach uses a form of inductive logical programming where causal relation extraction patterns are expressed as logical rules, which are processed using an inference engine | Mulkar-Mehta *et al.* (2011b); Mulkar-Mehta *et al.* (2011c) |

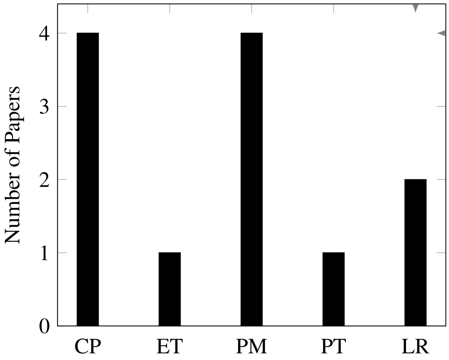


**Figure 1.** Popularity of research areas. CP, clue phrases; ET, extraction templates; PM, patterns matching; PT, parse trees; LR, logical rules.

Krishnan *et al.* 2014; Cao *et al.* 2015; Mirza 2016; Cao, Sun, and Zhuge 2016). These approaches adopt a number of common techniques which are summarised in Table 1, and whose popularity is shown in Figure 1.

The clue-based approach (Sakaji *et al.* 2008b; Ishii *et al.* 2010a; Cao *et al.* 2014; Mirza 2016) was one of the most frequent approaches within the linguistic approaches. Clue-based approaches rely upon lists of words and phrases which indicate the presence of a causal relation. Causal verbs such as ‘causes’ are common indicators of causal relations.

Clues can be hand-curated lists, or they can be part of a lexical resource. The former are typically small, which is highlighted by Sakaji *et al.* (2008b), who manually compiled a list of 40 clues. Clue-based approaches that rely upon lexical resources such as WordNet (Miller 1995) and VerbNet (Schuler 2005) are dependent upon predefined extraction patterns such as *NP V NP*, which produces causal relations such as: *Excessive sun causes sunburn*. The pattern represents cause and effect events as noun phrases (*NP*), and causal verbs that connect the events are represented by *V*. VerbNet is a representation of English verb behaviour (Palmer, Bonial, and Hwang 2017), and consequently, it is possible to replace the causal verb in a predefined extraction pattern with an alternate causal verb from VerbNet. For example, the verb ‘cause’ in the aforementioned phrase of *Excessive sun causes sunburn* could be replaced by the verb ‘provokes’ which is from the same grouping as ‘cause’ in VerbNet.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

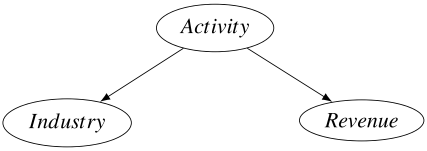


**Figure 2.** Example template (Onyshkevych 1993).

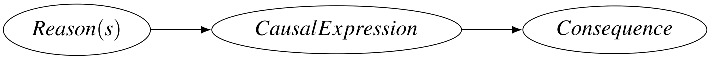


**Figure 3.** Example Seke template (Chan and Lam 2005).

Another resource that can be used in the clue-based approach is WordNet. This resource represents causation as a CAUSE-TO relation which 'is a transitive relation between verb synsets' (Girju and Moldovan 2002). The verbs that form part of the CAUSE-TO relations have 'nominalisation' (Girju and Moldovan 2002) from which it is possible to find causal relations between nouns. For example, WordNet defines a causal relationship between starvation and bonyness (Girju and Moldovan 2002).[d] From this causal pair of nouns, it is possible to identify further causal verbs from text using the pattern *NP V NP*, such as induces (starvation *induces* bonyness), where the verb *induces* is identified from its association with the nouns starvation and bonyness. The weakness of this technique is that it relies upon general lexical resources that may not represent causal connections comprehensively.

The template approach (Chan and Lam 2005) is based upon a general information extraction technique that uses a predefined template to extract relevant information from a document. An example of an information extraction template is shown in Figure 2. This is a hypothetical template proposed by Onyshkevych (1993) who wished to identify revenue activity phrases within a specific industry. Causal relations can be modelled similarly with a potential template that identifies cause and effect events connected by a causal bridge such as a verb or conjunction. This is an approach proposed by Chan and Lam (2005) and their Seke system.

The Seke system has seven templates that modelled causal relations extraction patterns. The extraction process is order dependent and the templates are ordered from left to right. An example template is shown in Figure 3 where the template identifies the reason, causal expression and the consequence. An example provided by the authors for a causal relation extracted by the aforementioned template is: *The increase of interest rates caused the Hang Seng Index to surge*. In this example, the reason is the *increase of increase rates* and the consequence is the *is the surge of the Hang Seng Index*.

The pattern-based approach for extracting causal relations relies upon the identification of frequently occurring patterns that identify causal relations. For example, in the commonly cited example of *Smoking causes cancer*, the pattern can be deciphered from part-of-speech (POS) tags, which are *noun, verb, noun*. The pattern can be generalised to other causal relations such as *rain causes floods* and *poverty creates violence*. Patterns may also be derived from dependency trees, and between phrases. For example, a common pattern between phrases for causal relation extraction is *noun-phrase, verb, noun-phrase* which can produce causal relations such as *lack of money creates poverty*.

The pattern-based approach for CRE was followed by several authors (Girju 2003; Cole *et al.* 2006; Ishii *et al.* 2010a; Sadek 2013). The research of Girju (2003) used a pattern-based approach to extract causal relations for a question-answering system. They used a general pattern of *NP Verb NP*. Their approach used the 429 causal relationships between nouns contained in WordNet. The

[d] http://wordnetweb.princeton.edu/perl/webwn?s=maceration.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

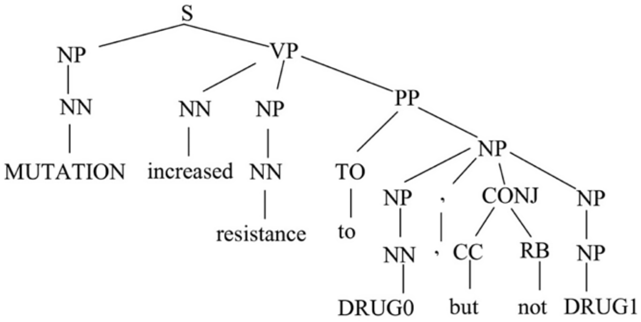


**Figure 4.** Generalised parse tree for medical causal relations (Bui *et al.* 2010).

nouns in the causal relationships are then used to extract verbs from a large collection of texts. In this way, they were able to extract further relationships between nouns from the document collection using the newly discovered verbs from the previous steps. In essence, this is a bootstrapping approach where a base set of relations is expanded using a pattern where the constituents of a causal relation are learnt from a large document collection. The approach produced sixty casual verbs and 681 causal patterns based on the *NP Verb NP* pattern. Causal relations extracted by this method were 'hunger causes headache' and 'movement triggers earthquake'. The causal relations extracted by this method tend to be simple because they follow the *NP Verb NP* pattern and would not be able to represent compound causal relations such as ones that have conjunctions such as 'and/or'.

Parse trees is a method for representing dependencies between words and phrases in a sentence in the form of a tree. The cause and effect relation has an inherent dependency where an effect cannot occur without a cause (Vendler 1967) and therefore can be represented in a parse tree. There was one work that followed this approach (Bui *et al.* 2010), which used causal relations from text to identify mutations in bacteria that cause drug resistance. A generalised example of the relationship that the authors wanted to identify is shown in Figure 4. The authors constructed eleven rules that extracted the desired relationship. An example rule is: *subject (keyword1) + Predicate (relation word + keyword2*), where *keyword1* and *keyword2* are MUTATION and DRUG, where MUTATION is one of a set of mutations, and DRUG is one of a set of drugs that the bacterial mutation will generate resistance to. The pattern requires three steps to extract a causal relation. The first step identifies the keyword pair, whereas the second step identifies *manner words* which are words that are present in a sentence and are at a distance of two or three words away from the keywords. The third step extracts the relation using information from steps one and two.

The logical rules approach is a form of inductive logic programming where a base set of logical rules is constructed and an inference engine performs inference on text using this base set of logical rules. This is the approach proposed by Mulkar-Mehta *et al.* (2011b); Mulkar-Mehta *et al.* (2011c). Mulkar-Mehta *et al.* (2011b) follow a typical approach. Their technique converts a sentence into its logical form and uses Mini-TACITUS which is an abductive inference engine (Mulkar-Mehta *et al.* 2011b) to perform inferences on the aforementioned logical form of a sentence using a set of axioms (Mulkar-Mehta *et al.* 2011b).

Linguistic approaches are relatively simple techniques that can use pre-existing lexical resources or manually identified clues as well as rules to identify causal relations. The advantage of these techniques is that they are relatively quick to develop and will be reasonably accurate.

A small sample of the results returned by linguistic approaches is shown in Table 2. The table demonstrates the strengths and weaknesses of this approach. The strength of the linguistic approach is that it does well in small corpora where the notion of causality is tightly defined, for example, the linguistic approach gained an F-Measure of 0.87 in Bui *et al.* (2010), where

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

**Table 2.** A sample of linguistic approaches

| Reference | Evaluation criteria | F-Measure | Corpus/sample size |
|---|---|---|---|
| Cao *et al.* (2014) | Manually selected verbs to collect causal relations from Google using the aforementioned verbs. The target language was Chinese | 0.92 | 5000 sentences |
| Bui *et al.* (2010) | Manually selected causal relations between mutation and drug. The 2937 candidate sentences gathered from articles in Pubmed each contained the aforementioned triple | 0.87 | 2937 sentences. 1913 were single sentences and 1024 were inter-sentences |
| Sakaji *et al.* (2008b) | A selection of 200 news stories about economic trends | 0.57 | 200 news stories |
| Cole *et al.* (2006) | A selection of 55 subject verb object triples | 0.74 | 55 triples (subject verb object) |

the domain was medicine and the subset of causation that the rules detected was the relationship between a mutation and a drug. However, in the broader domain of news where the causal relationship was less tightly defined, the rule-based approach obtained a much lower F-Measure (Sakaji *et al.* 2008b).

The literature review suggested that, outside of tightly defined notions of causation, this approach gains high precision, but at the expense of low recall. Thus, for broad definitions of causation in large corpora, linguistic approaches are unlikely to reach state-of-the-art results. This may be due to ambiguity in causal relations, where words with causal properties can have roles where no causation is present, and events can be both a cause and an effect. An example causal relation which would cause issues with a rule-based extraction system is: 'The causes are getting richer from donations from our sponsors'. This phrase will cause problems for a rule-based extraction system because the word 'causes' does not have causal properties, but a rule-based system will often assume that it will. And, the phrase 'from donations' is both a cause and effect, and a rule-based system will not be able to handle the dual role of this phrase. However, for small and niche applications, such as identifying causes for a specific disease, linguistic approaches are a viable technique.

### *3.2 Machine learning*

Machine learning is a series of techniques where an algorithm to can 'adapt to new circumstances and to detect and extrapolate patterns' (Russell, Norvig, and Norvig 2003) and is often referred to as learning from data.

There are three main types of machine learning techniques that are used in extracting causal relations from text. They are supervised learning, unsupervised learning and semi-supervised learning. Supervised learning for causal relations is where an Oracle or human expert labels causal relations in sentences from a document. The learner produces a model from the labelled data, and the model then delimits causal relations in unannotated text. Unsupervised techniques rely upon the similarity of causal relations, and therefore unsupervised techniques such as clustering group together causal relations from the surrounding text. Semi-supervised learning uses a mix of labelled and unlabelled data to produce a model from a learner.

An oversimplified image of the supervised learning process is shown in Figure 5. A table summarising the types of machine learning approaches used in CRE can be found in Table 3.

Each of the steps in Figure 5 can influence the performance of the final model. The preprocessing and annotation of the corpus were not described or did not vary between the papers that were discovered in the research literature. The learners and selected features varied between the

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

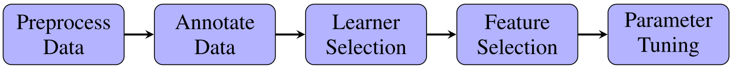


**Figure 5.** Simple representation of supervised learning phases.

papers. An evaluation of the learners found that random forest (Barik *et al.* 2017) and conditional random field (CRF) (Mihaila and Ananiadou 2013) were the most effective learners for CRE. The common features found in the literature review were unigrams, prefixes, suffixes and POS tags (Mehrabi *et al.* 2013; Riaz and Girju 2014). It should be noted that supervised machine learning strategies typically outperformed their linguistic cousins (Bhaskoro, Akbar, and Supangkat 2015).

The supervised techniques found in this paper are what could be called *classical machine learning*. In recent times, there has been a shift in the relation extraction research community to large neural networks and pre-training from large corpora as well as transformers (Wolf *et al.* 2019), which have produced state-of-the-art results, and these techniques are likely to produce superior results than the ones previously described. The literature review failed to provide a prescriptive method for estimating the complexity and the amount of labelled data required to extract causal relations. However, the literature review suggests that labelled data as low as a hundred examples may be sufficient to achieve the state-of-the-art results (Eisenschlos *et al.* 2019) for multilingual document classification, and 4750 sentences for CRE tasks (Li *et al.* 2021). Other factors may impact the accuracy of the model, in particular the definition of causal relations. If such relations span more than one sentence, they are likely to be more difficult to detect, because dependencies between words and phrases are at a longer distance than causal relations within a single sentence. Causal relations can have the cause phrase on the left- or right-hand side of the causal link, and it is likely that these dependencies can be captured by bidirectional recurrent neural networks (RNNs) (Zhou *et al.* 2016). The construction of a causal relation can add further complexity because an effect can be the cause of another causal relation. For example, 'the deforestation of the Amazon has caused a drought which has reduced the yield of sugarcane'. The drought is an effect of deforestation as well as the cause of the reduced yield of sugarcane. This dual role where an event can be a cause and an effect may cause issues with machine learning methods. The impact of the dual role of events is discussed in Section 3.5.

### *3.2.1 Embeddings and language models*

Classical machine learning represents words as tokens that are independent of their wider semantic representation such as synonymic and hyponymic relations. The relationships with other words are inferred through the training process. A weakness of this approach is that the relationships between words are limited to the training data. There have been several approaches from the general NLP community that uses unlabelled corpora to infer relationships between words. The two main approaches found in the literature review are word embeddings and language models.

Word embedding techniques such as Word2Vec (Mikolov *et al.* 2013) represent words as a vector that contains information about its co-occurrence within a fixed window with other words. The assumption behind this representation is that semantically similar words occur in similar contexts and would be represented by similar vectors.

Vectors are a coordinate system, consequently, a word vector can occupy a hypothetical location in a semantic space. This location can be compared with another word's location, and a similarity or distance value can be computed using measures such as cosine similarity. This allows supervised techniques that use this form of word representation to determine that the causal verb 'provoke' is similar to the causal verb 'incite', and dissimilar to the noun 'apple'.[e] Therefore, words that form a causal relation but do not explicitly appear in the training data can still be identified

[e] An example of word similarity calculations can be found here.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

**Table 3.** Summary of machine learning approaches

| Approach | Description | Citations |
|---|---|---|
| Classical supervised learning | This approach uses supervised learning and a non-neural network learner | Barik, Marsi, and Öztürk (2017), Mihaila and Ananiadou (2013), Riaz and Girju (2014), Son *et al.* (2017) |
| Supervised learning | This approach uses supervised learning and a neural network | Dasgupta *et al.* (2018), Li *et al.* (2019), Papanikolaou, Roberts, and Pierleoni (2019), Kyriakakis *et al.* (2019), Hassanzadeh *et al.* (2020), Yu, Li, and Wang 2019, Jin *et al.* (2020), Yang, Wu, and Hattori (2020a) |
| Semi-supervised learning | This approach uses a mix of labelled and unlabelled to produce a model which is then used to extract causal relations | Mihăilă and Ananiadou (2014), Drury and de Andrade Lopes (2015) |
| Unsupervised learning | This approach takes correlation approach to causal relations and assumes that events that co-occur frequently have a causal relationship | Riaz and Girju (2010), Riaz (2010), Do, Chan, and Roth (2011), Kim *et al.* (2012), Kim *et al.* (2013), Abinesh (2017), Hu, Rahimtoroghi, and Walker (2017) |

in unlabelled data because they are semantically similar to words and phrases that appear in the training data.

An example of this approach was proposed by Dasgupta *et al.* (2018). They used GloVe (Pennington, Socher, and Manning 2014), a method for word embedding, as well as long short-term memory (LSTM) (Hochreiter and Schmidhuber 1997), an architecture of RNN (Jain and Medsker 1999) that allows the estimation of long-distance dependencies in a sequence of information, for identifying causal clusters of related events. They evaluated their technique in SemEval-2010 Task 8[f] (Hendrickx *et al.* 2009), *Multi-Way Classification of Semantic Relations Between Pairs of Nominals*. An example of a cause cluster discovered by this technique is: *faulty fuel tank*, whereas an example effect cluster contained the phrase *risk of fire*.

A weakness of using word vectors generated with techniques such as Word2Vec or GloVe is that a word has a single vector, and therefore information about different word senses and contexts is compressed into a single representation.

This limitation may impede the extraction of causal relations because the context of a word, normally a verb, determines whether the surrounding NPs have a causal relationship. For example, the word ‘cause’ can be either a noun or a verb. The verb form of ‘cause’ is the one that can create a causal relationship between NPs. Newer forms of word embedding such as ELMo (Peters *et al.* 2018) and Flair (Akbik *et al.* 2019a) have multiple vectors which allow the embedding technique to be context-aware. This is the approach taken by Li *et al.* (2019), who developed the Self-Attentive BiLSTM-CRF wIth FlaIr Embeddings algorithm. The LSTM captures the long-distance dependencies of the causal relation, and the CRF computes the transition probabilities between the words. The Flair algorithm is used to compute a character language model from the 1-billion word benchmark corpus, which is used as an embedding layer in the LSTM. The technique was evaluated on the SemEval-2010 Task 8, which is a task to extract causal relations. The technique achieved an F1 Score of 0.85 on a test set of 804 sentences.

At their most basic language models estimate a probability over a given sequence of words. Initially, language models used statistical techniques to estimate the probability of the sequence. In the recent past, other techniques such as masked language modelling, where in the language model training phase portions of a sentence are removed, and the model attempts to predict the masked portion (Devlin *et al.* 2019). The modern language models are trained on large corpora

[f] https://www.kaggle.com/drtoshi/semeval2010-task-8-dataset.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

**Table 4.** Comparison of embeddings and language models for classifying causal sentences (Yang *et al.* 2020a)

| Technique | Test set | Precision | Recall | F1 score |
|---|---|---|---|---|
| Word2Vec | Wikipedia | 83.71 | 85.66 | 84.68 |
| | dialogue | 50.20 | 60.34 | 54.80 |
| BERT | Wikipedia | 94.36 | 94.80 | 94.58 |
| | dialogue | 65.41 | 70.23 | 67.73 |
| GPT-2 | Wikipedia | 91.24 | 90.34 | 90.79 |
| | dialogue | 62.22 | 64.55 | 63.36 |

and consequently, learn relationships between sequences of words. The ability to predict a likely sequence of words makes language models a good candidate technique for CRE.

Popular neural language models include XLNET (Yang *et al.* 2019b), BERT (Devlin *et al.* 2019) and GPT-2/3 (Radford *et al.* 2019), and it is likely that there will be differences in success between the language models when extracting causal relations because they are trained in different ways.

There has been an attempt to compare the ability of Word2Vec word embeddings, BERT and GPT-2 to detect causality in spoken dialogue (Yang *et al.* 2020a). The authors claim that they have a causality transfer learning technique which is a mix of fine-tuning and domain adaption. Although the authors are trying to identify causality in spoken dialogue, they used Wikipedia as the training medium because it is a rich source of causal statements. They used 64,658 sentences for the fine-tuning step which is an even split between causal and non-causal sentences. The domain adaption step is using the models trained on Wikipedia to evaluate a hundred causal sentences extracted from the Cornell Movie-Quotes Corpus (Danescu-Niculescu-Mizil and Lee 2011). The authors offered no discussion about pre-training; therefore, it can be assumed that the default distributions of these embeddings and language models were used. The embeddings and language models were included in a neural network, where a bi-directional gated recurrent unit (GRU) was used to classify the sentence into causal or non-causal categories. The results are shown in Table 4, and it is clear from the results that the domain adaption was relatively unsuccessful with Word2Vec scoring slightly better than chance. BERT and GPT-2 scored significantly higher than Word2Vec, and BERT scored marginally higher than GPT-2 in both tasks. However, the difference is marginal, and from the Wikipedia experiments, it seems that both language models with the GRU score highly and are suitable for causal sentence classification.

Language models and word vectors, both static and contextual, are often generated from large general corpora such as the Common Crawl (Common Crawl 2021) or Wikipedia. An alternative technique is to create specific causal embeddings from smaller corpora. This is an approach proposed by Sharp *et al.* (2016), who stated that causal embeddings capture 'complementary information' (Sharp *et al.* 2016) about the causal relationship between words that are missing from general word vectors. Causal embeddings are generated by extracting causal relations using a linguistic approach that has high precision, but low recall. These rules generate a triple: cause, causal connector and effect. This information is used as context for a word embedding technique. Sharp *et al.* (2016) report that this technique of word embedding, when combined with a convolutional neural network (CNN), outperformed general word embedding techniques with a CNN on a question-answering task. Causal embeddings are part of a wider trend, where specialised forms of word embeddings are used for specific tasks (Jastrzebski, Lesniak, and Czarnecki 2017).

### *3.2.2 Representation learning*

Representation learning is a subfield of machine learning which produces learners that can automatically find the representations of the data required for a specific task; therefore, tasks such as

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

feature selection are not required. A common representation learner is neural networks, and as shown in the previous section, neural networks have been successfully used in CRE tasks.

A drawback of using neural networks to extract causal relations is that they can require large amounts of labelled data to train a model from scratch. For example, there are training sets of over 500,000 labelled examples available for relation extraction tasks (Kumar 2017). Some strategies can be used to avoid labelling large amounts of data, such as transfer learning (Mou *et al.* 2016), which uses knowledge or relationships captured by another model in a different domain and adapts it to a new domain. As previously mentioned, this family of techniques requires a relatively small amount of data. There are several transfer learning strategies such as domain adaption (Ben-David *et al.* 2010; Sogaard 2013) and pre-training (Mikolov *et al.* 2018). Domain adaption is a technique that uses a model trained on a domain with a plentiful supply of labelled to domain with the same feature space, but with a different distribution (Ben-David *et al.* 2010). This is the approach taken by Ittoo and Bouma (2011a), who used Wikipedia as a base domain because it has numerous causal relations. They adapted a model trained on Wikipedia to extract causal relations from domains that are sparse in causal relations. Pre-training is a technique where the parameters of a neural network are learnt and fine-tuned for specific domains (Mikolov *et al.* 2018). Several approaches take that approach to CRE (Papanikolaou *et al.* 2019; Kyriakakis *et al.* 2019; Yu *et al.* 2019; Hassanzadeh *et al.* 2020; Jin *et al.* 2020; Yang *et al.* 2020a). This is the approach proposed by Kyriakakis *et al.* (2019), which is a representative paper of the area. They used a combination of BERT and a self-attention network (Vaswani *et al.* 2017) to classify whether a sentence from a biomedical corpus is either causal or non-causal. Their technique replaced the logic regression layer of the BERT neural network with the aforementioned self-attention network. They fine-tuned this layer with a small amount of labelled data. Their approach achieved an F1 score of 91.45 on the SemEval-2010 data set.

Supervised learning relies upon the availability of labelled data. In particular, manually labelled data can be expensive to obtain. An alternative to supervised learning is semi-supervised learning, where limited amounts of labelled data are used to train a model, and unlabelled data are then used to learn the marginal distribution of the data, which in turn improves the conditional distribution. This justification with the advent of transfer learning is now weaker because the amount of labelled data required for a supervised technique has been substantially reduced. The two located semi-supervised techniques predated the rise in popularity of language models and neural networks.

Mihăilă and Ananiadou (2014) used self-training, where a base learner is trained from labelled data. The model then classifies unlabelled instances, and then the high confidence classifications are added to the training data. These training data are then used to induce a new learner. The process continues until no more new training instances are found. Self-training is a relatively simple technique to implement but has a central flaw, which is errors in the training data or within the initial classification, cycles can be propagated throughout the training process. Despite this issue, Mihăilă and Ananiadou (2014) claimed that self-training improved its base learner. Co-training uses two or more separate views of the data source. Each view is used to train an individual model, and the models are used in a voting arrangement to assign a value to an unlabelled instance, and this instance can be used as training data. Co-training was used by Drury and de Andrade Lopes (2015) to extract causal relations from Portuguese news stories.

### *3.2.3 Unsupervised learning*

The CRE techniques discussed thus far have assumed that the author of a text is an Oracle who will explicitly describe a causal relation. The explicit description is encoded into a causal relation through the use of a causal verb, conjunction or continuation. This may not always be the case, and causal relations between events can be inferred through the co-occurrence of events within a fixed window, such as a sentence or paragraph (Riaz and Girju 2010; Riaz 2010). For example, the co-occurrence of the event bad weather and failing crops could be used to infer that bad weather

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

caused the failure of the crops. The problem with this approach is how to determine the cause and effect events. The aforementioned example does not provide any indication of which event is the cause and what event is the effect. Also, events can co-occur by chance, where there is no causal connection between the events. Any unsupervised technique will require some post-processing or filtering step to separate correlated events from causally connected events.

The advantage of identifying the co-occurrence of events is that no labelled data are required, and the techniques discovered in the literature review which use this technique are referred to as unsupervised learning.

Unsupervised techniques have used measures such as term distribution (Do *et al.* 2011), pointwise mutual information (PMI) (Do *et al.* 2011), Granger causation (Kim *et al.* 2012; Kim *et al.* 2013; Abinesh 2017) and causal potential (Hu *et al.* 2017) to identify frequently co-occurring events. A representative example of the unsupervised approach is the PMI technique which infers a causal connection through a PMI value between events. However, this technique will produce a small number of causal relations.

### *3.2.4 Hybrid techniques*

Machine learning and linguistic CRE strategies can be combined to form a hybrid technique. These techniques use linguistic knowledge to bootstrap labelled data or filter data to use in a supervised learning technique (Blanco, Castell, and Moldovan 2008; Ittoo and Bouma 2011b; Krishnan *et al.* 2014; Cao *et al.* 2015; Mirza and Tonelli 2016; Hidey and McKeown 2016; Dasgupta *et al.* 2018). A common strategy is to choose corpora that are likely to have a high concentration of causal patterns such as Wikipedia or parallel corpora (Hidey and McKeown 2016). In common with linguistic CRE strategies, hybrid strategies have used hand-coded lists of causal patterns or causal indicators from external lexical resources as a source of causal patterns (Cao *et al.* 2012).

An outlier in the methods surveyed for this survey articles was by Ning *et al.* (2018), who proposed that temporal and causal relations are bound to each other with one relation dictating the result of the other. They proposed an integer linear programming technique that enforced constraints inherent in time and causality. The authors claimed by approaching temporal and CRE as a joint problem that their technique improves existing relation extraction strategies.

Another outlier in the casual relation techniques is Liu, Chen, and Zhao (2020) who uses background information from ConceptNet, an ontology, as well as a reasoner, to supplement information from the texts that are being used to extract causal relations. The authors use mask-based training where events are removed from texts so that the learner can learn contextual information around events. And finally, an attention mechanism coordinates the first two phases to identify causal relations between events. This technique scored a F-score of 45.40 on the event causality data set released by Do *et al.* (2011). In comparison, the PMI technique scored 23.30 (Do *et al.* 2011). Two further techniques: ECD PMI (Do *et al.* 2011) and CEA (Do *et al.* 2011), scored 29.90 and 38.60. It is clear on Do *et al.* (2011)'s data set that the strategy proposed by Liu et al. (2020) is superior.

### *3.2.5 Discussion*

The dominant approach in machine learning causal relation is supervised learning. With the advent of language models and transformers (Wolf *et al.* 2019) that can detect long-distance dependencies within a sentence, this process should be accelerated. This acceleration is shown in Figure 6 where the type of learner used and the article publication year is shown for the papers discovered in this section. It is quite clear that since 2018, neural networks have been the learner of choice for CRE.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

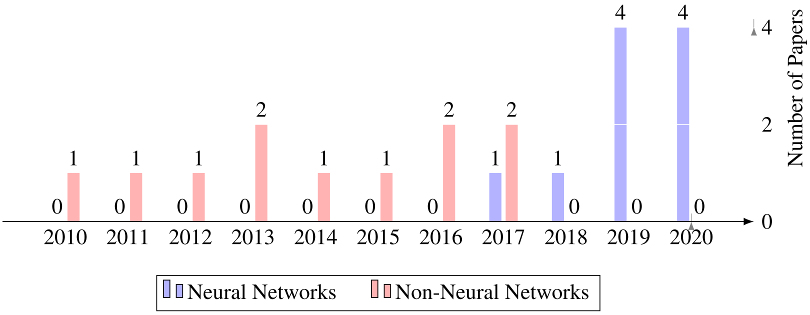


**Figure 6.** Popularity of learners over time.

### 3.3 Corpora for CRE

CRE strategies often require collections of labelled text documents to train machine learning strategies and evaluate machine learning, and linguistic approaches. The research literature revealed that there are a number of labelled corpora that can be used to evaluate and train CRE strategies.

The resources that were discovered in the literature review were 'BECauSE' (Dunietz, Levin, and Carbonell 2017), 'TempEval-3' (Mirza *et al.* 2014), 'Richer Event Description (RED)' (O'Gorman, Wright-Bettner, and Palmer 2016), Balanced Corpus of Contemporary Written Japanese (Kaneko and Bekki 2014a), Japanese Corpus for Causal Relations (Kaneko and Bekki 2014b), FinCausal (Mariko *et al.* 2020) and the Event Storyline Corpus (Caselli and Vossen 2017). The BECauSE corpus annotation schema contains a cause, a causal connective and an effect, and the causal relation is sub-divided into TEMPORAL, HYPOTHETICAL and CORRELATION categories. The corpus is relatively small as it has 59 randomly selected news articles (Dunietz *et al.* 2017).

The TempEval-3 (English) (Mirza *et al.* 2014) and RED (English) (O'Gorman *et al.* 2016) were not originally annotated for causal expressions, but they have had their annotation extended to encapsulate causal information. TempEval-3 corpus initially had time expression, event and temporal information annotated (UzZaman *et al.* 2013). Mirza *et al.* (2014) extended the corpus annotation scheme with causal information using a rule-based algorithm. The algorithm annotates a causal relation that may have one of the following categories: CAUSE, ENABLE and PREVENT. The RED corpus is relatively small with 95 documents and has a similar annotation scheme to the one proposed by UzZaman *et al.* (2013), but with the conditional, and causes information that is captured with the following tags: BEFORE/CAUSES, OVERLAP/CAUSES, BEFORE/PRECONDITION and OVERLAP/PRECONDITION. The corpora described thus far have had differing annotation schemes, and there have been other annotation schemes suggested for causal information. For example, Dunietz, Levin, and Carbonell (2015) proposed: CONSEQUENCE, MOTIVATION, PURPOSE and INFERENCE annotations, whereas Mirza and Tonelli (2014), Mostafazadeh *et al.* (2016) have extended TimeML with the CAUSE, ENABLE and PREVENT tags.

As demonstrated by the summary of the corpora in Table 5, the majority of corpora discovered in this literature review were small and they had fine-grained causal annotations, which were variants of CAUSE, ENABLE and PREVENT. Each of these annotation schemes does not provide a comprehensive view of causation, which may bias any CRE strategy based upon any of the individual corpora discovered in the literature review. The relatively small size of the corpora would have caused problems for modern neural networks, but the advent of transfer learning will mitigate the small size of the available corpora.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

**Table 5.** Summary of the characteristics of a sample of causal corpora

| Corpus name | Size | Tags | Citations |
|---|---|---|---|
| BECauSE | 59 documents | TEMPORAL, HYPOTHETICAL, CORRELATION | Dunietz *et al.* (2017) |
| TempEval-3 | 500K tokens of silver standard data. 100K tokens of gold standard data | CAUSE, ENABLE and PREVENT | Mirza and Tonelli (2016), Mirza (2016), Ponti and Korhonen (2017) |
| RED | 95 documents | BEFORE/CAUSES, OVERLAP/CAUSES, BEFORE/PRECONDITION, and OVERLAP/PRECONDITION | Li *et al.* (2019) |
| Balanced Corpus of Contemporary Written Japanese | 66 sentences | CAUSAL, NO-REL | Maekawa *et al.* (2014) |
| Japanese Corpus for Causal Relations | 128 sentences | Alternation, Consequence, Contrast, Elaboration, Explanation, Commentary, Instance Addition, Parallel, Narration, Introduction and Background | Kaneko and Bekki (2014a) |
| Fin Causal | 3405 documents | Cause, QFact, Fact and Discard/Remove | Mariko *et al.* (2020) |
| Event Storyline Corpus | 258 documents | Action_Occurrence, Action_State, Action_Reporting, Action_Aspectual, Action_Perception, Action_Causative and Action_Generic | Kyriakakis *et al.* (2019) |

### 3.4 Counterfactual CRE

This section thus far has discussed the extraction of causal relations that contain causatives that bridge cause and effect events with approaches that can be described as correlations where one event correlates with another. In the brief discussion about causality, other forms of causation such as counterfactuals were discussed. These techniques tend to be outliers in the literature and are not used in the applications of causal relations.

There are a small number of papers that describe resources and techniques for detecting counterfactuals (Son *et al.* 2017; Ojha *et al.* 2020; Akl, Mariko, and Labidurie 2020; Nwaike and Jiao 2020; Yang *et al.* 2020b; Fajcik *et al.* 2020). All but one paper found in the literature review is based around the SemEval-2020 Task 5 (Yang *et al.* 2020b) which had two tasks. The first task was a classification task which identified if a sentence contained a counterfactual statement, and the second was an extraction task which identified the ‘antecedent and consequent’ (Yang *et al.* 2020b) in a counterfactual. The technique proposed by Son *et al.* (2017) used a combination of rules and a support vector machine (SVM) to capture sentences with counterfactual statements. A dependency parser is used to extract the arguments from the counterfactual sentence. The technique extracted a subset of counterfactual causal relations. They are documented in Table 6.

### 3.5 Causal chains, dependent and independent causes

Causation in text is often not a discrete cause and effect, and therefore chains can be created where an event can be a cause and effect. These events can play this dual role because the event is not a root cause or a terminal effect. None of the papers surveyed tried to detect chains or disambiguate them. Any attempt at causal chain detection will need labels outside cause and effect that indicate that the event has a dual role. The disambiguation may have to be part of a supervised learning technique rather than post-processing a set of events, because the dual role may impact the accuracy of a model that uses cause and effect labels.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

**Table 6.** Summary of counterfactual extraction (Son *et al.* 2017)

| Counterfactual form | Example |
|---|---|
| Wish Verb | I wish I had been richer |
| Conjunctive Normal | If everyone put differences aside and get along, everything would be so much enjoyable |
| Conjunctive Converse | I would be stronger, if I had lifted weights |
| Modal Normal | They should of shown this guy getting shot, that would have been TV gold. |
| Verb Inversion Had | If I left the event early, I would not have met John |
| Should Have | I should have joined the event early |
| (Would/Could) Have | I would have been happier without John |

Dependent causes are causes that rely upon each other, and the absence of one will ensure that the effect event will not trigger. Independent causes are causes that occur together, but only one is sufficient to trigger the effect event. This is a form of both early and late preemption. This type of information can be determined using conjunctions such as ‘and’/‘or’. These conjunctions link events, and with this information, it is possible to determine if the cause events are dependent or independent.

The papers surveyed for this paper did not have any papers that tried to identify more than one cause per effect event, and consequently, there are no works referred to in this survey that tried to identify dependent or independent causes. It is likely that the identification of dependent causes will improve reasoning systems built on causal relations.

### 3.6 *Implicit, explicit and inter-sentential causation*

Thus far, the article has described techniques that have extracted explicit causal relations. Explicit causal relations are often referred to as marked causation. There is another type of causation which is implicit causation which can be referred to as unmarked causation. The paragraph that discussed unsupervised learning describes some techniques that can detect implicit causal statements. There are, however, other techniques that can be used to find implicit causal statements.

Inter-sentential causation is where cause and effect are separated by punctuation and therefore the cause and effect may appear in different sentences (Jin *et al.* 2020). Examples of the differences between intra-sentential and inter-sentential as well explicit and implicit causation are given by Jin *et al.* (2020).

In common with explicit causation, there are a number of methods that can extract inter-sentential (Wu, Yu, and Chang 2012; Jin *et al.* 2020) and implicit causation (Ittoo and Bouma 2011b; Ittoo and Bouma 2011a; Grivaz 2012; Kilicoglu 2016; Jin *et al.* 2020).

The work conducted by Jin *et al.* (2020) addressed both implicit and inter-sentential causation. Their approach used a supervised learning approach and a novel neural network learner which they called cascaded multi-structure neural network (CSNN). The CSNN has three phases and combines two learners and a self-attention network. The first phase is a CNN which captures local features from separate sentences. These features are correlated using a self-attention network. And finally, a BiLSTM is used to determine the long-distance dependencies. The authors provide several examples that their model returns, for example, ‘affected by the market the gross profit fell’. The target domain that they performed their experiments on was Chinese financial texts. They claim an F1-measure of 0.82 for causal extraction and 0.75 for effect extraction.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

Despite the novel architecture, this approach will still be limited by the common issues suffered by machine learning methods that were highlighted earlier. The main issues are labelled data dictating the type of causation returned, the exploding gradient problem, as well as determining the difference between correlated events and cause/effect events. In essence, this is still a supervised learning approach, with all its inherent shortcomings. This area however, is a start of a greater exploration of causation in text than the common triple of NP, causal verb, NP, and is likely to yield more informative causal information.

### 3.7 Causal explanation detection

This is a subtask of CRE which are explanations of why an event or action has happened. This is a relatively small area of causality and therefore there were a limited number of papers discovered in the literature review. The dominant approach is machine learning (Son, Bayas, and Schwartz 2018;Zuo *et al*. 2020b; Zuo *et al*. 2020a). For example, Zuo *et al*. (2020b) used pyramid salient-aware networks. The technique identifies keywords in a discourse structure, and from these keywords identify the root or subject of the discourse. The dependencies of the root are also identified. From the same sentence, a causal explanation is found. The aforementioned work used a self-attention network to identify the causal explanation. The technique score was 0.87, 0.77 and 0.82 on the Facebook, PDTB-CED and BECauSE-CED data sets, respectively (Zuo *et al*. 2020b). The BECauSE-CED data and PDTB-CED sets are modified versions of the BECauSE (Dunietz *et al*. 2017) and Penn Discourse TreeBank (PDTB) (Rashmi *et al*. 2008) corpora. The PDTB-CED assumed that 'causal semantic discourse relations' (Zuo *et al*. 2020b) were treated 'as supplemental messages with causal explanations' (Zuo *et al*. 2020b), whereas BECauSE-CED assumed that 'sentences containing causal span pairs in BECauSE Corpus' were also treated 'as supplemental messages with causal explanations'.

### 3.8 Emotion cause identification

The survey thus far has discussed CRE as an information extraction task that identifies a causal relationship between events. This survey argues that causal relations in natural language require cause and effect phrases, and therefore emotion cause detection is a specialised form of CRE where the cause of an emotion is identified. An example of emotion cause is shown in the following statement: *He was excited because he received a promotion at work*, where the sentiment in the phrase is represented by the word excited, which is caused by the event promotion at work. There is a subtle difference between classical event-based CRE and emotion cause detection; however, CRE techniques can be adapted to identify the causes of emotion because the relationship between sentiment phrases and the cause event is similar to the relationship between cause and effect events.

The emotion cause identification extraction literature is similar to that of causal relations where there are two distinct approaches to extract emotion causes which are rules (Lee, Chen, and Huang 2010; Chen *et al*. 2010; Lee *et al*. 2013) and machine learning (Li and Xu 2014; Xu *et al*. 2017; Chen *et al*. 2020).

#### 3.8.1 Rule-based emotion cause identification

Rule-based approaches rely upon linguistic rules which are either learnt from a corpus using rule learning algorithms, such as association rule learning or more commonly manually constructed rules. Manually constructed rules can reply upon POS tags or dependency parsing of the text. The domain expert then infers a relationship between the emotion and the cause event.

There were three rule-based approaches discovered in the literature review (Lee *et al*. 2010; Chen *et al*. 2010; Lee *et al*. 2013). The rule-based techniques are typified by the approach followed

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

by Lee *et al.* (2010). They used linguistic indicators such as 'the position of cause event, verbs, epistemic markers, conjunctions, and prepositions' (Lee *et al.* 2010) to create rules that identify emotion causes. They classified the type of emotion into one of the following groups: happiness, sadness, fear, anger and surprise, that their rules could identify their causes.

The rules constructed by Lee *et al.* (2010) are dependent upon causal verbs such as causes or provokes to identify the causal relationship between the emotion and its cause. The strategy that was followed by Lee *et al.* (2010) identified a sentence with a sentiment or emotion keyword. This sentence will be referred to as the 'focus sentence'. They also capture the preceding sentence (prefix sentence) of the focus sentence and the following sentence (suffix sentence).

The authors produced 15 general rules, which, when evaluated, produced 47.95 F-score.

Rule-based approaches for information extraction in general, and emotion cause identification especially, should be avoided because manual rule writing is a labour-intensive process and only captures a small sample of emotion causes, and updating rules takes a significant amount of time.

### *3.8.2 Machine learning-based emotion cause identification*

The two machine learning techniques for emotion cause identification discovered in the literature search were supervised learning techniques. Supervised techniques rely upon labelled data where labels are attached by a domain expert to words and phrases and the learner uses this information to produce a model. The model then infers labels on unseen information.

The first approach proposed by Li and Xu (2014) used an SVM to identify the cause event rather than the whole relation. They used linguistic markers as features. Their technique was evaluated on manually annotated posts from the Chinese social media site, Weibo. Their system achieved an F-score of 61.30 which is better than the rule-based system by Lee *et al.* (2010).

The second approach proposed by Xu *et al.* (2017) used an ensemble approach where multiple learners work in cooperation to label unseen data as a cause event for sentiment in a sentence. The learner chosen in this technique was again an SVM. The authors define an event as: action, actor, object. The authors claim that the action is the verb and the actor is the subject of a phrase. The ensemble approach uses different kernels for SVM in a bagging approach where each learner is given a subset of the labelled data. The data are prepossessed by a dependency parser so it is possible to present to the models, a tuple containing a causal event candidate. They claim that this technique has a 7.00% gain over the state of the art.

The emotion cause identification techniques are similar to the older CRE strategies, and therefore these techniques can likely be greatly improved upon by the adoption of modern techniques that use RNN, attention networks and pre-training such as BERT.

### *3.8.3 Corpora for emotion cause identification*

There are several linguistic resources for emotion cause identification that can be used for training and evaluation. For example, the 'Up to Now' corpus (Gui *et al.* 2014) is an emotion cause corpus for Chinese. The annotated text is taken from the Chinese micro-blogging Weibo. Social media is not the only media that can be used to build corpora for emotion cause. Gui *et al.* (2016), for example, used Chinese news to build an annotated corpus which used the EmotionML[g] annotation schema.

The corpora discovered in the literature review were exclusively for Chinese and were relatively small-in-size. This limitation is likely to inhibit future development of emotion cause identification techniques.

[g] https://www.w3.org/TR/emotionml/.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

**Table 7.** Summary of emotion cause identification techniques

| Technique | Citations | Summary |
|---|---|---|
| Rule-based | Lee *et al.* (2010) | Manually constructed rules that identify the cause of emotion. The rules use often use triggers or keywords to identify the causes |
| Clue-based approach | Chen *et al.* (2010), Lee *et al.* (2013) | This technique used linguistic clues in the form of rules to identify causes of emotion |
| Supervised learning (SVM) | Li and Xu (2014) | An SVM trained on manually annotated posts to find emotion causes |
| Ensemble learning | Xu *et al.* (2017) | Bagging approach where each learner takes a sub-sample of the annotated training data. A voting approach is used to identify emotion causes |

#### *3.8.4 Emotion cause discussion*

As shown in Table 7, the literature for emotion cause detection is less rich than for CRE; however, the discovered techniques were similar to that of a CRE. Due to the similarity of the two domains, it is likely that the emotion causes domain will follow the CRE domain where sequence classifiers such as LSTM, GRU and transformers will begin to dominate future research in this area. In common with the CRE literature, there is a lack of freely available corpora, and the resources that were found were mainly designed for the Chinese language. This will for the immediate future limit the adoption of the use of modern neural networks to identify emotion causes.

### ***3.9 Discussion***

The advent of large document collections and the adaption of neural networks for sequence classification as well as the development of embeddings have ensured that the current state of the art for CRE has evolved from linguistic and CRF-based techniques to LSTM and RNNs with word vectors.

Linguistic methods are limited to patterns and the words identified manually or present within lexical resources, therefore the causal relations that are returned will be limited to these patterns. Causal relations whose patterns fall outside the manually constructed linguistic rules or contain clues or words outside those in lexical resources will not be captured. This will lead to a very limited set of causal relations being returned by linguistic methods.

Supervised learning methods can generalise causal relations from labelled data. The main issue for classical machine learning techniques such as CRF is:

- They do not capture long-distance dependencies between cause and effect because of limitations of the learner.

The more recent variations of RNN such as LSTMs and GRUs have extended the ability to capture long-distance dependencies. However, there are issues with these types of learner, they are:

- They suffer from the vanishing or exploding gradient for long sequences where cause and effect events are a long distance apart.
- The type of causal relation returned is dictated by the labelled data.

Despite the drawbacks, supervised learning is likely to remain one of the most popular CRE techniques.

Unsupervised methods that rely upon correlations between events are unlikely to progress without considering the nature of causal relations as demonstrated by Weber, Rudinger, and Van

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

**Table 8.** Popularity of research areas

| Research area | Frequency |
|---|---|
| Causal information modelling | 14 |
| Event prediction | 10 |
| Cause identification | 8 |
| Question-answering | 8 |
| Text summarisation | 7 |
| Reasoning | 4 |
| Information retrieval | 3 |
| Sentiment analysis | 3 |

Durme (2020). This paper used the notion of manipulation in determining the temporal order of correlated events to form a causal relation or chain. The use of these techniques is not as informative as the supervised methods because the supervised methods will detect the nature of the causal relationship between cause and effect through resultative and instrumental causal verbs.

At the time of writing, it is reasonable to assume that the advances in CRE will follow improvements in long-distance dependency detection from the neural network research community as well as the increasing size of language models.

## 4. Applications of CRE

Causal relations frame a causal relationship between two or more events. This relationship implies that the existence of one event will force the occurrence of one or more events. This unique relationship allows for the exploitation of the information to produce unique or improve existing applications of NLP.

There are several applications of causal relations discovered in the literature review, and the break-down of these applications can be found in Table 8. The common applications are causal information modelling, event prediction, cause identification, text summarisation and information retrieval (IR).

Causal information modelling is a modelling technique that aggregates related causal relations into a graph or related structure. Event prediction is a technique that predicts future events based upon their causal relationships with current or past events. Cause identification is a technique that traces the causes of current events. Question-answering is an application that allows the posing of questions in natural language which will be answered by the system in natural language. Text summarisation is a strategy of reducing long pieces of text into a shorter structure that outlines the main or important parts of the longer text. Reasoning is a process where new information is inferred from base axioms. IR is a strategy of returning relevant documents in response to a query from a document collection. Finally, sentiment analysis is a technique of identifying opinionated language. It is clear from Table 8 that the most frequent areas of research are causal information modelling and cause identification.

### *4.1 Causal information modelling*

The representation and the consequent visualisation of causal relationships within a specific domain can aid the understanding of causal influences as well as the prediction of future events.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

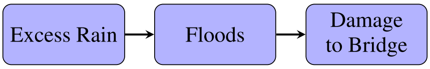


**Figure 7.** Simple causal chain.

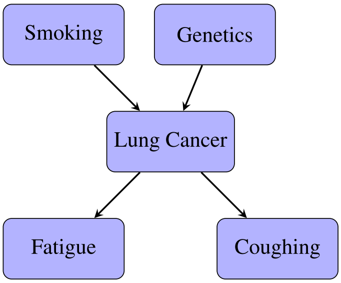


**Figure 8.** Simple causal network.

The role of causal relations in this task is to provide a network of causally linked events from which inferences or explanations about events can be made. The causal relations are extracted from a corpus of relevant documents and are aggregated into a network of events.

Causal relations are typically a one-to-one relationship between cause and effect events. Conjunctions in the cause or effect event can alter the causal relationship to one of the following: many-to-many, many-to-one or one-to-many. The relationship between the multiple events in the cause phrase may be independent, for example, *rain or snow caused problems with the trains*, or dependent, *snow and rain caused problems with the trains*. In each case, the causal effect is immediate, the snow or the rain caused problems with the trains. The aggregation of causal relations can form a causal chain where an initial cause acts through one or more agents to create an effect event. An example of a causal chain is shown in Figure 7, where the initial cause *Excess Rain* acts through the agent *Floods* to force the effect event, *damage to bridge*. This form of simple causal information modelling illustrates causes of causes and may assist an end-user to understand the flow of causality in a domain.

A causal network extends the idea of a causal chain by creating a directed graph that represents the causal connections between events (Jensen 2001). Figure 8 shows a simple oft-cited graph of the factors that cause lung cancer. Unlike the causal chain, the causal network shows multiple independent causes of an effect event, as well as multiple effects.

There were a number of papers that constructed causal networks (Ishii, Ma, and Yoshikawa 2010b; Ishii *et al.* 2010a; Ackerman 2012; Puente, Garrido, and Olivas 2013a; Puente, Olivas, and Prado 2014; Luo *et al.* 2016; Zhao *et al.* 2017; Kang *et al.* 2017) from causal relations extracted from documents. These networks were constructed with the purpose of supporting other tasks such as news understanding (Ishii *et al.* 2010a), text summarisation (Puente *et al.* 2014), question-answering (Puente *et al.* 2013a) and commonsense reasoning (Luo *et al.* 2016).

The news understanding causal network created by Ishii *et al.* (2010a) was a topic-event casual network where cause and event events are linked together. The explanation of a news event was supplied using event and topic keywords on each of the nodes. The event keywords are given context by the topic keywords which complement the causal description provided by the edges. The text summarisation approach proposed by Puente *et al.* (2014) constructed a causal graph from the extraction of causal relations from the target document. A summary generated by the causal network is extracted by identifying the most relevant information using standard extractive summarisation techniques such as term-frequency inverse document frequency (TF-IDF). The order of the final summary is dictated by the flow of causality in the causal network. Question-answering

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

systems are applications that users pose questions to, and receive answers in natural language. Causal information can assist question-answering systems, by identifying causes and effects. An example question whose answer depends upon causal information is, *What was the cause of the second world war?*. The approach proposed by Puente *et al.* (2013a) used a causal network as background information to support a question-answering system. The approach was almost identical to their technique for extractive summarisation (Puente *et al.* 2014). Commonsense causal reasoning 'is the process of capturing and understanding the causal dependencies amongst events and actions' (Luo *et al.* 2016). The causal network provides background information to assist causal inferences between words and phrases.

A semantic network 'is a graph structure for representing knowledge in patterns of interconnected nodes and arcs' (Shapiro 1992). A semantic network is a more general representation of relationships between concepts than a causal network because cause and effect in this representation is one type of relationship. Semantic networks can be undirected; however, the ones that contain causal information must be directed. Semantic networks can also be referred to as knowledge graphs. This type of representation was used as background information by Cao, Sun, and Zhuge (2018) to assist with the summarisation of scientific papers.

The graphs that have been described thus far are a form of a knowledge graph, which is an undirected graph with edges and nodes, with the nodes representing an entity and the edge representing a relationship between the entities.

There are a number of articles that use the term knowledge graph, and construct it using causal relations. This technique has been used in the health (Rotmensch *et al.* 2017; Bakal *et al.* 2018; Yu 2020a), chatbots (Yu 2020b) and drug (Sastre *et al.* 2020) domains. The papers discovered in this review found articles describing symptoms and the disease that they are associated with (Rotmensch *et al.* 2017). The construction of the knowledge graph for such a domain would require entity disambiguation, that is, there would need to be a standard way of describing diseases and their symptoms. And because the causal relations in Rotmensch *et al.* (2017) were drawn from health records, there would need to be a correction mechanism for errors in the text as well as errors in recording symptoms and disease diagnosis. This application of causal relations is an interesting domain because it may form part of an automated diagnosis system that can identify diseases in humans without human intervention.

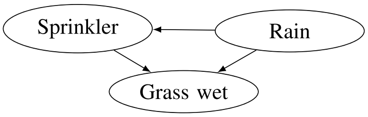
A richer representation of causal information is a Bayesian network. Bayesian networks are probabilistic directed graphs that consist of random variables (nodes) and edges between nodes that can represent a causal relationship between one or more parental node, and a child node. The relationship between the parental nodes and the child node is represented in a conditional probability table (CPT) which has a combination of the parental states and the probability of a child node state. A simple Bayesian network is shown in Figure 9, which is a well-known example of computing the probability of the grass is wet based on the likelihood of the sprinkler being switched on or the probability of it raining.

The advantage that Bayesian networks have over causal networks is that not only are Bayesian networks an illustration of the causative relationships in a domain but because each node has a CPT, it is possible to estimate the likelihood of an event occurring given the occurrence of one or more cause events.

Bayesian networks can be created from causal relations extracted from text. In the literature review, there were found two approaches that use causal relations to build Bayesian networks (Sanchez-Graillet and Poesio 2004; Drury *et al.* 2016). The approach described by Drury *et al.* (2016) is a typical example where causal relations were extracted with a semi-supervised approach (Drury *et al.* 2016) and the events were transformed into nodes and the nodes were connected if they co-occurred in a causal relation. The states of the nodes were one of the following, positive, negative or neutral, and the probabilities of these states were computed by the frequency of the sentiment states of the parental node with states of the child nodes. Sentiment states were computed by identifying sentiment words in the causal relations. For example, 'Strong winds causes

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

|  | Sprinkler |  |
|---|---|---|
| Rain | T | F |
| F | 0.4 | 0.6 |
| T | 0.01 | 0.99 |



| Rain |  |
|---|---|
| T | F |
| 0.2 | 0.8 |

|  |  | Grass wet |  |
|---|---|---|---|
| Sprinkler | rain | F | T |
| F | F | 0.4 | 0.6 |
| F | T | 0.01 | 0.99 |
| T | F | 0.01 | 0.99 |
| T | T | 0.01 | 0.99 |

**Figure 9.** Simple Bayesian network (Murphy 2012).

devastation to the harvest' would have the relationship 'neutral' to 'negative' because the cause event 'Strong winds' has no sentiment words and the effect event has the negative sentiment word devastation. The paper provided strategies for breaking cycles, which cannot be modelled in static Bayesian networks, and coarsening which reduces the number of nodes in the network. The reduction of nodes is an important step because Bayesian networks constructed from text will be sparse and because reasoning in Bayesian networks is computational complex, it may not be possible to reason with these networks. Also, the paper describes a technique for reducing the number of parental nodes which in turn limits the size of the CPTs in the network. Large CPTs may not be possible to compute in a reasonable period because the growth of the CPT is exponential which is represented by $a^{n+1,]}$, where $a$ is the number of nodes states and $n$ is the number of parental nodes for a given child node. This paper shows the practical challenges of constructing a knowledge representation from text where the informality of the representation of causal information in text can cause the creation of large sparse graphs. Because of this informality compromises will need to be taken to create a usable Bayesian network.

Causal information modelling has several applications which were described in this subsection. Aggregating causal relations into a representation such as a causal network can give a detailed overview of a domain. There is a drawback to this approach, for example, the aggregation of causal relations can create very large, but very sparse graphs. However, with pruning and merging techniques, it is possible to remove some of the sparseness.

### *4.2 Event prediction*

Event prediction, which is also known as script event prediction or script inference, is a task where a future event is inferred based upon causal information in current events or chain of events. The role of causal relations in event prediction is to either: provide the effect events of potential cause events in news stories or related documents, or be aggregated into the aforementioned causal chain.

As previously discussed, cause and effect events have a causal relationship where the occurrence of the cause event will force the appearance of an effect event. An example of this is demonstrated by the following relationship: *Downturn in the economy causes large scale unemployment*. From the sentence, it can be inferred that, in the case of a downturn in the economy event, a large-scale unemployment event will occur at a specific time in the future.

Event prediction is a relatively popular area of research (Sakai and Masuyama 2007; Sakaji, Sakai, and Masuyama 2008a; Radinsky, Davidovich, and Markovitch 2012; Kunneman and van

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

den Bosch 2012; Radinsky and Horvitz 2013; Hashimoto *et al.* 2014; Preethi *et al.* 2015; Hashimoto *et al.* 2015; Pichotta and Mooney 2016; Li, Ding, and Liu 2018; Yang *et al.* 2019a) which has been used to predict general events (Kunneman and van den Bosch 2012; Radinsky *et al.* 2012; Radinsky and Horvitz 2013; Hashimoto *et al.* 2014; Hashimoto *et al.* 2015; Preethi *et al.* 2015) as well as to indicate future economic/business trends (Sakai and Masuyama 2007). The common approaches for event prediction are causal patterns (Radinsky *et al.* 2012; Preethi *et al.* 2015; Hashimoto *et al.* 2014), semantic patterns (Hashimoto *et al.* 2015), graph neural networks (Li *et al.* 2018; Yang *et al.* 2019a) and LSTM/RNN (Pichotta and Mooney 2016; Li *et al.* 2018).

The approaches to event prediction match that of the CRE where the classical approaches rely upon pattern /linguistic-based approaches, whereas the more modern approaches use neural networks in combination with word embeddings or language models.

The most frequent classical technique used for event prediction using causal information from text is the causal pattern approach. A causal pattern approach extracts causal pairs from text through linguistic patterns, such as the previously mentioned: *NP V NP*. These causal relations can be aggregated into a causal chain, from which future events can be predicted depending on the position of a current event is in the causal chain. An example of this approach is described by Hashimoto *et al.* (2014), who generated causal chains of linked events from causal relations extracted from text. Their system was developed to identify the effects of social problem events such as deforestation. Their approach identified a base set of social problem events (NPs) extracted from Wikipedia. These events were connected using causal relations into event chains where two or more events are connected in a linear form to create a chain. In total, more than two million event chains were generated. A large number of those described the same series of events, but the textual description was slightly different. For example, synonyms of the same concept, such as omit and neglect, would have produced a new causal chain. The authors culled approximately 1.2 million causal chains because they replicated existing causal chains. Their event prediction method used the causal chains to predict hypothetical future events from the existence of the original cause phrase. A scenario generated from this approach was: 'deforestation continues → global warming worsens → sea temperatures rise → vibrio parahaemolyticus fouls (water)' (Hashimoto *et al.* 2014). The authors claimed that this series of events was never described in a single document, and based upon events described in documents in 2007 they were able to predict an event that was recorded in 2013.

A similar approach to the causal pattern is semantic relation extraction, which seems to take the same approach as causal patterns where causal relations are extracted from a corpus. However, because a semantic approach identifies not only causal relations between events but also between the nouns of the cause and effect event, it is possible to infer new causal relations through the substitution of subject nouns with nouns that have a semantically similar relationship. This is the approach proposed by Hashimoto *et al.* (2015). They claim that using their approach they can infer the causal relationship 'deploy a security camera to avoid crimes' (Hashimoto *et al.* 2015) from 'deploy a mosquito net to avoid malaria' (Hashimoto *et al.* 2015) through the semantic relationship 'A PREVENTS B' (Hashimoto *et al.* 2015).

The causal chain or causal relation approach would seem to be limited because it assumes a one to one causal relationship where one event forces the effect event. This, however, may ignore the combination of events causing an effect event. The aforementioned causal modelling approach[h] allows the estimation of the effect of multiple cause events. The approach proposed by Drury *et al.* (2016) created a Bayesian network from information in agricultural news, which can be used to estimate the impact of news events upon the agricultural market.

The more recent approaches have relied upon variations of RNN or graph neural networks. For example, Pichotta and Mooney (2016) used an LSTM to learn the dependencies within text at the sentence level and predict a missing event in a hold-out evaluation. They trained their system on

[h] See Page 44.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

Wikipedia, which is a rich source of causal relations, and held-out around 1% of the data gathered from Wikipedia. They found that the model that was trained at the sentence level produced better results than LTSMs trained at the event level.

#### *4.2.1 Discussion*

Event prediction is not only an interesting research area but has practical real-world applications. For example, the prediction of events can assist stock-traders who would have an advantage over their peers because event prediction will give them a larger trading horizon (Ding *et al.* 2014; Izumi and Sakaji 2019). Also, disaster management (VanVactor 2010; Qiu *et al.* 2017) is an applicable area, where knowledge of the chain of future events will be an advantage because preemptive action can be undertaken once a cause event has occurred.

The current approaches of event prediction rely upon chains of events that are inferred from co-occurrence within a specific window. It is likely that script inference and event prediction tasks will begin to draw from the philosophy of causation to produce more nuanced techniques that do not rely upon correlations of events within fixed windows.

### *4.3 Cause identification*

Post-hoc analysis of events is often required for tasks such as disaster investigations. For this survey, we will refer to this task as cause identification. Cause identification is the inverse of event prediction where the technique takes a current event, such as extreme weather or unpopularity of a product, and identifies its causes, such as climate change or poor build quality. This application also has some overlap with knowledge modelling where causal networks are used to assist in the explanation of an event (Ishii *et al.* 2010a; Ishii *et al.* 2010b; Ackerman 2012; Ackerman 2013; Zhao *et al.* 2017).

The role of causal relations is similar to that of event prediction where causal relations extracted from text are used to build causal networks or a related model from which root cause or causes of events can be identified. The causal relations for this task are extracted from a corpus and are aggregated into a causal network.

The dominant themes of this area were the identification of causes of accidents (Tirunagari *et al.* 2012; Sizov and ÖztÜrk 2013) and machine failure (Joskowicz, Ksiezyck, and Grishman 1989). Cause identification for accidents and machine failure can allow for the identification of common themes which then can be eliminated or mitigated to reduce the future incidence of accidents or machine failure.

A representative paper from this area is Tirunagari *et al.* (2012) who used causal relations extracted from ‘MAIB accident investigation reports’ to find the root causes of marine accidents. The corpus contained 135 documents. The reports contained eleven types of accidents, but the authors concentrated on four types: collisions, groundings, fire and machine failure. From the causal relations extracted from this type of accidents the authors manually created concepts that classified the cause and effect event. For example, the word ‘abuse’ was part of the labour-relations concept, and the word diesel was mapped to the concept of fuel. The frequency of the concepts was displayed graphically, and therefore this graphical representation will allow investigators to identify the common causes of accidents.

Cause identification is a niche application of causal relations with few papers discovered in the literature review. However, these techniques perform an explanatory function where current events can be explained with a graphical representation of the causal chain or a combination of events that have created the current event under investigation. In addition to the discovered papers about accidents and machine failure, it is possible that other domains such as politics and international relations could use cause identification techniques to aid the understanding of current events.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

### *4.4 Question-answering*

Question-answering is a subfield of IR where humans can pose questions to a system and receive a reply in natural language (Kolomiyets and Moens 2011). Causal information can assist in question-answering strategies for specific questions where the question asks for a cause or a consequence. The literature review revealed a small number of question-answering systems (Girju and Moldovan 2002; Girju 2003; Pechsiri and Kawtrakul 2007; Higashinaka and Isozaki 2008a; Higashinaka and Isozaki 2008b; Oh *et al.* 2013; Puente *et al.* 2013a; Sharp *et al.* 2016) that use causal relations to increase the relevance of the returned answers.

The role of causal relations in question-answering is to assist in the formulation of an answer to a specific type of question. Effect events and phrases tend to appear in answer candidates, and therefore questions that have cause phrases and events can be answered with sentences and phrases which contain effect events or phrases.

Also, causal relations extracted from text may assist question-answering to answer specific types of questions. 'Why Questions' such as 'Why was X arrested?', where the hypothetical answer could be 'X was arrested for fraud' (Higashinaka and Isozaki 2008a), are suited to systems that use causal relations in their technique (Pechsiri and Kawtrakul 2007; Higashinaka and Isozaki 2008a; Higashinaka and Isozaki 2008b; Oh *et al.* 2013). The use of causal relations in question-answering systems produces highly relevant answers to 'Why Questions' (Higashinaka and Isozaki 2008a).

In addition to 'Why Questions', causal relations in question-answering systems can be used to respond to 'Causal Questions' (Girju 2003; Girju and Moldovan 2002), such as 'What are the causes of lung cancer?' (Girju and Moldovan 2002) and 'Name the effects of radiation on health' (Girju and Moldovan 2002).

It is unlikely that a system will rely exclusively upon information in causal relations, and therefore the main role of causal relations is to assist more general approaches such as general lexical models (Sharp *et al.* 2016). Also, causal relations may help in the answer construction process (Puente *et al.* 2013a).

Question-answering techniques have benefited through the addition of information extracted from causal relations because causal relations contain a cause event and an effect event. Systems that are wholly dependent upon causal relations are unlikely to be complete systems that can provide answers to general questions. However, the literature review found that causal relations can assist in the answering of 'Why' and 'Causal' questions.

### *4.5 Text summarisation*

Text summarisation is a process where a single document or multiple documents are summarised into a single shorter textual description which represents the main themes of the source document or documents. Causal relationships can be used to improve existing techniques or create novel summarisation strategies by ensuring that the cause and effect are in the correct order in the summary.

The literature review found four types of summarisation that have used textual causal relations. The strategies covered structured summarisation (Zhang, Jatowt, and Tanaka 2016a, 2016b), abstractive summarisation (Puente *et al.* 2014; Puente *et al.* 2013b; Puente *et al.* 2016) and extractive summarisation (Puente *et al.* 2017).

The role of the causal relations in these tasks depends upon the type of summarisation. There are however three main roles for causal relations: ensuring that events are summarised in temporal order, an indicator of a high information sentence or act as a form of knowledge representation which can be used to aid the summarisation process. Causal relations can guide the temporal order of summarisation because the relationships encoded into a causal relation or a causal chain have a temporal order where an effect cannot occur before the cause event. Therefore in a summarisation process, the causal relation will prevent the event ordering from being out of order. Causal relations can be a source of information because it encodes a causal relationship between events,

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

and therefore sentences that contain causal relations are likely to be informative as the causal relation contains an explanation. For example, 'Due to the formula explicitly outlined in the law, the tax rate on gasoline and diesel fuel will increase on Oct. 1',[i] where an explanation for the tax rises is provided in the cause event.

Structured summarisation is a strategy that generates a timeline of temporally ordered events that represents the source document or documents (Binh Tran 2013). The main theme of the structured summarisation was that of product evolution overtime (Zhang *et al.* 2016a; Zhang *et al.* 2016b). A typical example is an approach proposed by Zhang *et al.* (2016a) who used causal relations extracted from the Amazon Product Review Data set.[j] Their technique relies upon detecting causal relationships in the review text, and detecting changes when the cause term changes in the frequency of publication. These changes are summarised by a time-series graph.

The abstractive approach to text summarisation produces a substantially shorter text from concepts discovered from the source text(s). The summary text is not copied from the source text(s) but is rewritten in a manner to represent the main themes and concepts of the sources. The approach proposed by Puente *et al.* (2013b) constructed a causal network from causal relations in the text, and from the causal relationships in the graph, the text is rewritten using SimpleNLG[k] which is a generative grammar tool.

The extractive summarisation approach is a technique that uses the text from the source document(s) to produce a summary. This is achieved by ranking sentences so that the most informative are ranked higher than less information-dense sentences. Sentences that contain causal information often are an indicator of their ability to summarise a text. The extractive approach by Puente *et al.* (2017) was similar to several alternative abstractive approaches, but the causal graph was used to rank sentences that appear in the summary. There are other approaches (Li *et al.* 2006) that do not wholly rely upon causal information but use it as part of a larger strategy.

Summarisation utilises the relationships within causal relations to produce a summary of larger source texts. Causal relations can provide an overview of the connections between events in a document; however, it is unlikely that causal information alone can produce summaries and it is likely that causal information will form a part of a larger summarisation technique.

### 4.6 Reasoning

Reasoning from text is one of the oldest forms of logical inference and is referred to as syllogistic reasoning, where a conclusion is deduced from a minimum of two propositions. The most famous syllogism has two propositions 'all men are mortal' and 'Socrates is a man', from which the conclusion that 'Socrates is mortal' can be drawn. Reasoning with causal relations would require a cause as one proposition, and the conclusion as an effect. For example, one hypothetical causal syllogism could be 'smoking causes cancer', 'Socrates is a smoker', and the consequence would be 'Socrates will acquire cancer'.

The role of causal relations in this type of application is to represent the propositions of a reasoning system. In the aforementioned example of smoking causes cancer a causal relation will represent smoking as the cause of cancer. The causal relations are extracted from a relevant corpus and can be used as the basis of a reasoning system. It should be noted that causal relations alone are unlikely to be sufficient for a reasoning system, and other propositions are likely to be needed.

The literature review found that causal relations are likely to enrich the reasoning process rather than being the main focus. One example of a technique that uses causal relations to enrich a reasoning process was proposed by Ovchinnikova *et al.* (2014) who built an abductive reasoning (Aliseda 2006) system. Abductive reasoning seeks the most likely explanation given a set of

[i] https://patch.com/new-jersey/pointpleasant/nj-gas-tax-increase-due-coronavirus-crisis-officials.

[j] https://s3.amazonaws.com/amazon-reviews-pds/readme.html.

[k] https://github.com/simplenlg/simplenlg.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

evidence. And in their system, the authors used lexical resources such as WordNet and FrameNet as well as English Slot Grammar to convert text into a logical representation (Ovchinnikova *et al.* 2014). Causal relations in FrameNet were used to assist with the conversion of the text into a logical representation. The use of FrameNet and the causal relationships it holds as a basis for a reasoning system was also proposed by Ovchinnikova *et al.* (2010) who extended FrameNet to improve its effectiveness in its reasoning strategies.

Causal relations have its role to play in the choice of plausible alternatives (CPAs), which is a common application of causal reasoning. CPAs take a premise and offers two or more explanations, which the CPA technique will select as the most plausible explanation. For example, the following shows an example of CPA where there is one reasonable explanation and another which is not.

- Premise: The man broke his toe. What was the CAUSE of this?
- Alternative 1: He got a hole in his sock.
- Alternative 2: He dropped a hammer on his foot.[1]

There is a public evaluation of CPA techniques, SemEval-2012 task 7 (Gordon, Kozareva, and Roemmele 2012), which provides an open data set against which several systems could be evaluated. In this task, a premise is given and then a choice of two English sentences, one of which is the cause event of the initial sentences. An example given by Gordon *et al.* (2012) is as follows:

- Premise: The man fell unconscious
- Alternative 1: The assailant struck the man on the head.
- Alternative 2: The assailant took the man's wallet (Gordon *et al.* 2012).

The winning systems (Gordon *et al.* 2012) in this task used a PMI-based system similar to Do *et al.* (2011). The PMI-based systems selected the alternative with the highest PMI to the premise.

The final method found in the literature review used causal relations to construct a causal network which in turn was used to assist the reasoning process. This type of reasoning technique was covered on Page 44

Causal reasoning with text is a fringe research activity, and this may be due to the lack of annotated data and corpora. With suitable corpora, there is an argument that domains such as medicine, agriculture and finance would benefit from causal reasoning. The referred techniques not only can provide explanations of events and phenomena but also can discover new knowledge.

### *4.7 Information retrieval*

IR is a technique where users can enter keyword queries and the IR system returns a series of documents ranked by their similarity to the initial query. Well-known examples of IR systems are web search engines, which, in addition to similarity measures such as TF-IDF, use authority measures such as Page Rank (Page *et al.* 1999).

The role of causal relations in IR is typically in the word expansion, where a word can be associated with other words and phrases through the aggregation of causal relations. For example, the causal relation: 'cigarettes causes lung cancer', the words, cigarettes, lung and cancer would be associated with each other. This association can be used in query expansion or in the manner in which documents are returned by the proximity in sentences of causally related words.

There were a small number of techniques (Khoo 1996; Khoo, Myaeng, and Oddy 2001; Agueda 2010) discovered that used causal relations to improve the precision of IR systems. The IR systems that were discovered in the literature review were not wholly dependent upon causal relations, and

[1] http://people.ict.usc.edu/ gordon/copa.html.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

they were used in combination with techniques such as word proximity to improve the precision of the system (Khoo *et al.* 2001).

There was a lack of activity in the IR domain for a number of years; however, in 2020 there was new life breathed into the area with the causality-driven ad hoc information retrieval (CAIR) task (Datta *et al.* 2020b). This task released several documents that had causally related themes, and the task evaluated how well search systems found the causally related documents (Datta *et al.* 2020b). A related work is Datta *et al.* (2020a) which is similar to the CAIR task where causes are retrieved for the effect expressed in the query. The recent activity in this area is for causal retrieval rather than assisting traditional IR systems. The lack of progress in assisting traditional IR systems could infer that the IR community has concluded that causal approaches to IR are a research 'dead-end'.

### *4.8 Sentiment analysis*

Sentiment analysis is a technique to detect subjectivity in textual sources. Subjectivity in mass-media textual sources such as social and traditional media can be used to infer the public mood about stocks, shares, products as well as political figures. Because of this characteristic of mass-media, sentiment analysis has been used in trading strategies on the stock market (Li *et al.* 2014).

Techniques that use sentiment analysis to trade but only have access to news published on the Internet or through traditional sources have a disadvantage because there is a lag between the time news is published on a commercial system like a Reuters terminal, and the time the same stories appear on an Internet news site. Consequently, traders who use traditional sentiment analysis and online news are unlikely to be able to compete with traders who have access to proprietary news and use the aforementioned sentiment techniques. An alternative is to use casual–sentimental relations that have a cause event that creates an effect event that has a sentiment orientation (Dehkharghani *et al.* 2014). An example of casual–sentimental relation is *Excess rain causes a poor harvest*, where the negative effect event is indicated by the sentiment word 'poor'. The cause event *Excess rain* can be used to infer a future sentiment distribution (Drury and de Andrade Lopes 2015). Therefore, the cause event can be used to trade because it is an indicator of future sentiment. Future sentiment may allow a longer trading horizon and negate the lag of news emerging from proprietary information providers onto the internet.

Although casual–sentimental relations have not been directly used for stock trading, their potential has been hinted at by Preethi *et al.* (2015) and Drury *et al.* (2016). Preethi *et al.* (2015) used casual–sentimental relations to improve event predictions and their future sentiment orientation. A technique proposed by Drury *et al.* (2016) followed a similar route where events (nodes) and the interaction of sentiment states of each node were modelled in a Bayesian network. The Bayesian network could then be used to predict events and their future sentiment orientation.

Although there were a small number of papers found in this area, it remains an interesting area of research because it allows the inference of future sentiment distribution rather than the current sentiment distribution. And therefore this predictive capability should extend the horizon for inferences made upon sentiment information.

### *4.9 Applications and CRE methods*

The applications of causal relations are dependent upon the technique used to extract them, and therefore the method chosen will determine the nature of the application. Also the corpus that the causal relations are drawn from will also influence the application. Table 9 shows an overview of the applications with the type of CRE method that was used in the application. It is clear from the table that the most popular extraction method used by applications that were found in the

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

**Table 9.** Causal relation extraction methods used by applications

| Causal relation extraction method | Techniques |
|---|---|
| Causal cues | Luo *et al.* (2016), Ovchinnikova *et al.* (2014), Ovchinnikova *et al.* (2010), Puente *et al.* (2013b), Khoo *et al.* (2001), Girju and Moldovan (2002), Ishii *et al.* (2010b), Ishii *et al.* (2010a), Ackerman (2012), Sakaji *et al.* (2008a), Ackerman (2013), Pechsiri and Kawtrakul (2007), Oh *et al.* (2013) |
| Causality connectors | Zhao *et al.* (2017), Sanchez-Graillet and Poesio (2004) |
| Clustering | Radinsky and Horvitz (2013) |
| Co-cause (Zhang *et al.* 2016a) | Zhang *et al.* (2016a); Zhang *et al.* (2016b) |
| Discourse parser | Sizov and Öztürk (2013), Puente *et al.* (2016) |
| Extraction template | Hashimoto *et al.* (2014) |
| Keywords | Puente *et al.* (2014); Puente *et al.* (2013a) |
| Multi-view learning | Drury *et al.* (2016) |
| Patterns | Girju (2003), Sharp *et al.* (2016), Higashinaka and Isozaki (2008a), Sakai and Masuyama (2007), Tirunagari *et al.* (2012), Preethi *et al.* (2015), Cao *et al.* (2018) |
| Rule learning | Radinsky *et al.* (2012) |
| Semantic patterns | Hashimoto *et al.* (2015) |
| Supervised learning | Pichotta and Mooney (2016); Yang *et al.* (2019a); Li *et al.* (2018); Higashinaka and Isozaki (2008b); Kunneman and van den Bosch (2012) |
| Word proximity (Khoo 1996) | Khoo (1996) |

literature review was the causal clue technique. A similar technique, patterns, was the second most frequent. Machine learning was a relatively unpopular technique, however with libraries such as Hugging Face[m] and Flairs (Akbik *et al.* 2019b) that make the use of language models such as BERT less complex is likely to increase the uptake of supervised methods to build applications. It should be noted that two papers (Khoo 1996; Zhang *et al.* 2016a) created novel CRE techniques for their applications.

The breakdown of corpora used in the applications of causal relations is shown in Figure 10. It is clear from the chart is that the main corpora used are news and general corpora. Corpora are considered general if they are large and have a mixture of text document types. This domination by general and news corpora is due to applications that wish to resolve general problems such as reasoning, or support applications that analyse or infer events from news. There were a number of specialist corpora such as accidents and answers, which reflect the specific nature of the derived applications such as to cause identification (accidents) and question-answering (answers).

The domination of general and news corpora may not last because of the advent of transfer learning where small specialist corpora can be used to fine-tune large language models, which in turn will be able to extract causal relations from the aforementioned specialist corpora. This feature of large neural networks may in turn increase the popularity of supervised machine learning techniques to extract causal relations as well as the development of niche applications.

[m] https://huggingface.co.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

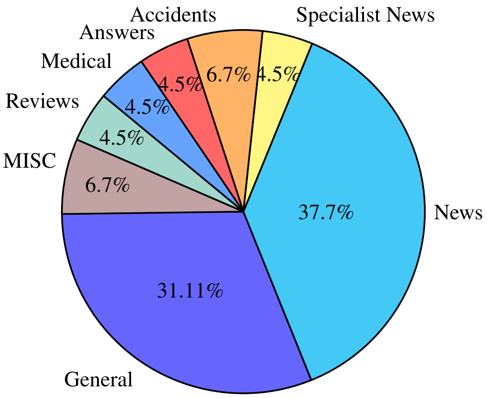


**Figure 10.** Breakdown of corpora used in causal relation applications.

### 4.10 Industry applications

The research literature has presented some possible opportunities for larger-scale applications for industry. The most obvious application is the modelling of causal relations into a form of background knowledge such as a knowledge graph or causal network which then can be used to inform applications such as automated disease diagnosis system, trading systems and drug side effects and interactions. There are several companies that are using these techniques to produce novel applications because large text collections will have several non-obvious causal connections which can be discovered through causal chains. Automated systems then can infer the impact of a combination of drugs or the impact of an event on the stock market.

Event prediction can offer industry a way of predicting the outcome of decisions and government policies. Large-scale news analysis using event prediction and graphs may allow for a better understanding of foreign and domestic policy.

Finally, causal relations could be used to verify false claims in advertising such as health (Yu *et al.* 2020). False advertising claims can be a laborious task to prove, and therefore an automated system could be used to expedite false advertising legal processes.

Causal relations extracted from text offer industry an opportunity to create applications that are difficult or not possible with current or traditional approaches.

## 5. Conclusion

This survey intends to demonstrate not only the applications of causal relations but also the techniques available to extract causal relations and emotion causes as well as the labelled corpora.

The applications of causal relations span a wide number of domains, and because of the relationship between cause and effect event, causal relations can improve existing strategies. In several applications such as event prediction, causal relations are an important source of information that is unlikely to be captured in other information sources. These specific applications of causal relations are likely to have other applications in areas such as stock trading because of their ability to infer future events based upon current knowledge.

This paper argues that some applications that are potentially suitable for a causal relation approach are not documented in the research literature. These applications would be able to use causal relations to reason about a domain and infer new knowledge that is absent in a single document in the training literature. It is not possible to provide an authoritative list of applications, however, based on the information in the research literature it is a prediction of this paper that applications such as estimating side effects of drug combination therapy, and public policy

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

simulations would be good candidates. It should be noted that there are limitations to applications built upon causal relations. Applications that are likely to be successful will be able to use inductive reasoning, where the future repeats the past. This will tend to favour physical systems rather than social ones. In social systems, the agents will know the past and will react differently to the same stimuli in the future. Also, systems, where information is private, are unlikely to be successful.

The paper has not only discussed the applications of causal relations but how to determine the applicability of a domain as well. An applicable domain will have source literature that is causal relation dense, and these causal relations are consistent and do not vary over time. There are no standard data sets that are suitable for all domains, but, likely, using a language model (pre-trained) on large corpora will be a common starting point for causal relations application. These models will use a transfer learning technique such as fine-tuning or domain adaption on specialist corpora which will need to be collected for each project as the publicly available corpora are too small and specialised. The experiments that need to be run are similar to that of generic machine learning projects where hold-out or 10-fold cross-evaluation techniques are suitable to predict the robustness of a model to predict unlabelled examples. The validation of the final application will depend on the domain and application. The papers referred to in this survey provide examples of application and domain-specific evaluations.

The applications of causal relations are an underused technique, and therefore there may be areas that are not covered by this review, that could benefit from using causal information. It is hoped that this article can stimulate the use of causal information in new areas where causal relations are currently not used.

Any improvement of existing applications will be predicated upon the availability of causal relations. CRE strategies until recently have been simplistic, however with the advent of Transformers with large language models such as BERT (Devlin *et al.* 2019), and more recently MT-DNN (Liu *et al.* 2019), as well as transfer learning from models trained on complementary domains the modern techniques are likely to have higher precision and recall than their antecedents. The predominance of the use of Wikipedia in pre-training strategies is likely to assist causal relation strategies as demonstrated by some of the classical CRE strategies that use Wikipedia (Hidey and McKeown 2016).

Causal relations can improve existing strategies, but it is an assertion of this survey that the future direction of research should concentrate on the unique applications of the properties of causal relations such as event prediction, knowledge representation and reasoning. The availability of large corpora and cheap computing power should allow the construction of large models that provide a detailed representation of a domain, and these large models should be able to improve reasoning systems.

**Acknowledgements.** Brett Drury was partially funded by the State of São Paulo Research Foundation (FAPESP), Brazil by grant number https://bv.fapesp.br/pt/auxilios/98395/modelos-de-risco-para-a-agricultura-a-partir-de-informacao-em-texto/ 16/15524-3. Hugo Gonçalo Oliveira was partially supported by national funds through FCT, within the scope of the project CISUC (UID/CEC/00326/2020) and by European Social Fund, through the Regional Operational Program Centro 2020. Alneu de Andrade Lopes was partially funded by the State of São Paulo Research Foundation (FAPESP), Brazil, and the Brazilian National Council for Scientific and Technological Development (CNPq) grant 304040/2019-3. We would also like to acknowledge the role of the three anonymous referees who assisted the construction of this article with their suggestions and criticisms.

## References

**Abinesh S.** (2017). *Prediction of Future News Trends*. Master's Thesis, National University of Ireland Galway, NUIG, University Road, Galway, Ireland.

**Ackerman E.J.M.** (2012). Extracting a causal network of news topics. In **Herrero P.**, **Panetto H.**, **Meersman R.** and **Dillon T.** (eds), *On the Move to Meaningful Internet Systems: OTM 2012 Workshops*. Lecture Notes in Computer Science, vol. 7567. Springer, pp. 33–42.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

**Ackerman E.J.M.** (2013). *Extracting Causal Relations between News Topics from Distributed Sources*. PhD Thesis, Technische Universität Dresden.

**Adams F.C.** (2008). Stars in other universes: stellar structure with different fundamental constants. *Journal of Cosmology and Astroparticle Physics* **2008**(08), 010.

**Agueda C.P.** (2010). *Extraction and Analysis of Conditional and Causal Sentences for Information Retrieval*. PhD Thesis, Universidad Pontificia Comillas.

**Akbik A.**, **Bergmann T.**, **Blythe D.**, **Rasul K.**, **Schweter S.** and **Vollgraf R.** (2019a). FLAIR: an easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Minneapolis, Minnesota. Association for Computational Linguistics, pp. 54–59.

**Akbik A.**, **Bergmann T.**, **Blythe D.**, **Rasul K.**, **Schweter S.** and **Vollgraf R.** (2019b). Flair: an easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 54–59.

**Akl H.A.**, **Mariko D.** and **Labidurie E.** (2020). Semeval-2020 task 5: detecting counterfactuals by disambiguation. arXiv preprint arXiv:2005.08519.

**Aliseda A.** (2006). *Abductive Reasoning*, vol. **330**. Berlin, Germany: Springer.

**Altenberg B.** (1984). Causal linking in spoken and written English. *Studia Linguistica* **38**(1), 20–69.

**Bakal G.**, **Talari P.**, **Kakani E.V.** and **Kavuluru R.** (2018). Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations. *Journal of Biomedical Informatics* **82**, 189–199.

**Barik B.**, **Marsi E.** and **Öztürk P.** (2017). *Extracting Causal Relations Among Complex Events in Natural Science Literature*. Cham: Springer International Publishing, pp. 131–137.

**Barrow J.D.** and **Silk J.** (1980). The structure of the early universe. *Scientific American* **242**(4), 118–129.

**Beebee H.**, **Hitchcock C.** and **Menzies P.** (2015). *The Oxford Handbook of Causation*. OUP.

**Ben-David S.**, **Blitzer J.**, **Crammer K.**, **Kulesza A.**, **Pereira F.** and **Vaughan J.W.** (2010). A theory of learning from different domains. *Machine Learning*, **79**(1–2), 151–175.

**Bhaskoro S.B.**, **Akbar S.** and **Supangkat S.H.** (2015). Identification of causal pattern using opinion analysis in Indonesian medical texts. In *2015 International Conference on Information Technology Systems and Innovation (ICITSI)*. IEEE, pp. 1–7.

**Binh Tran G.** (2013). Structured summarization for news events. In *Proceedings of the 22nd International Conference on World Wide Web*. ACM, pp. 343–348.

**Blanco E.**, **Castell N.** and **Moldovan D.I.** (2008). Causal relation extraction. In *LREC*.

**Bui Q.-C.**, **Nualláin B.Ó.**, **Boucher C.A.** and **Sloot P.M.** (2010). Extracting causal relations on HIV drug resistance from literature. *BMC Bioinformatics* 11(1), 101.

**Cao M.**, **Sun X.** and **Zhuge H.** (2016). The role of cause-effect link within scientific paper. In *2016 12th International Conference on Semantics, Knowledge and Grids (SKG)*. IEEE, pp. 32–39.

**Cao M.**, **Sun X.** and **Zhuge H.** (2018). The contribution of cause-effect link to representing the core of scientific paper—the role of semantic link network. *PloS One* **13**(6), e0199303.

**Cao Y.**, **Cao C.**, **Zhang J.** and **Niu, W.** (2015). Two-phased event causality acquisition: coupling the boundary identification and argument identification approaches. In *International Conference on Knowledge Science, Engineering and Management*. Springer, pp. 588–599.

**Cao Y.**, **Zhang P.**, **Guo J.** and **Guo L.** (2014). Mining large-scale event knowledge from web text. *Procedia Computer Science* 29(0), 478–487. 2014 International Conference on Computational Science.

**Cao Y.-N.**, **Cao C.**, **Wang S.** and **Zang L.** (2012). Web mining for causal relations between events. *Information* **15**(1), 427–434.

**Casati R.** and **Varzi A.** (2020). Events. In **Zalta**, **E.N.** (ed.), *The Stanford Encyclopedia of Philosophy*, summer 2020 Edn. Metaphysics Research Lab, Stanford University.

**Caselli T.** and **Vossen P.** (2017). The event storyline corpus: a new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pp. 77–86.

**Chan K.** and **Lam W.** (2005). Extracting causation knowledge from natural language texts. *The International Journal of Intelligent Systems* **20**(3), 327–358.

**Chen Y.**, **Hou W.**, **Li S.**, **Wu C.** and **Zhang X.** (2020). End-to-end emotion-cause pair extraction with graph convolutional network. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 198–207.

**Chen Y.**, **Lee S.Y.M.**, **Li S.** and **Huang C.-R.** (2010). Emotion cause detection with linguistic constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, pp. 179–187.

**Cole S.**, **Royal M.**, **Valtorta M.**, **Huhns M.** and **Bowles J.** (2006). A lightweight tool for automatically extracting causal relationships from text. In SoutheastCon, 2006. Proceedings of the IEEE, pp. 125–129.

**Collins J.D.**, **Hall E.J.** and **Paul L.A.** (2004). *Causation and Counterfactuals*. MIT Press.

**Common** Crawl (2021). Common Crawl. Available at http://commoncrawl.org/ (accessed 22 March 2021).

**Copley B.** and **Martine F.** (eds.) (2015). Causation in Grammatical Structures, vol. 1. OUP.

**Copley B.** and **Wolf P.** (2015). Theories of Causation should inform linguistic theory and vice versa. In *Causation in Grammatical Structures*, vol. **1**. OUP.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

**Dai Z.**, **Yang Z.**, **Yang Y.**, **Cohen W.W.**, **Carbonell J.**, **Le Q.V.** and **Salakhutdinov R.** (2019). Transformer-xl: attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860.

**Danescu-Niculescu-Mizil C.** and **Lee L.** (2011). Chameleons in imagined conversations: a new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*.

**Dasgupta T.**, **Saha R.**, **Dey L.** and **Naskar A.** (2018). Automatic extraction of causal relations from text using linguistically informed deep neural networks. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 306–316.

**Datta S.**, **Ganguly D.**, **Roy D.**, **Bonin F.**, **Jochim C.** and **Mitra M.** (2020a). Retrieving potential causes from a query event. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1689–1692.

**Datta S.**, **Ganguly D.**, **Roy D.**, **Greene D.**, **Jochim C.** and **Bonin F.** (2020b). Overview of the causality-driven adhoc information retrieval (cair) task at fire-2020. In Forum for Information Retrieval Evaluation, pp. 14–17.

**Degand L.** (1994). Towards an account of causation in a multilingual text generation system. In *Proceedings of the Seventh International Workshop on Natural Language Generation*, INLG'94, pp. 108–116.

**Degand L.** (2000). Causal connectives or causal prepositions? discursive constraints. *Journal of Pragmatics* **32**(6), 687–707.

**Dehkharghani R.**, **Mercan H.**, **Javeed A.** and **Saygin Y.** (2014). Sentimental causal rule discovery from twitter. *Expert Systems with Applications* **41**(10), 4950–4958.

**Devlin J.**, **Chang M.-W.**, **Lee K.** and **Toutanova K.** (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics, pp. 4171–4186.

**de Spinoza B.** (1996). *The Ethics*. Penguin.

**Ding X.**, **Zhang Y.**, **Liu T.** and **Duan J.** (2014). Using structured events to predict stock price movement: an empirical investigation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1415–1425.

**Do Q.X.**, **Chan Y.S.** and **Roth D.** (2011). Minimally supervised event causality identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'11, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 294–303.

**Drury B.** and **de Andrade Lopes A.** (2015). The identification of indicators of sentiment using a multi-view self-training algorithm. *Oslo Studies in Language* **7**(1), 379–395.

**Drury B.**, **Rocha C.**, **Moura M.-F.** and **de Andrade Lopes A.** (2016). The extraction from news stories a causal topic centred Bayesian graph for sugarcane. In *Proceedings of the 20th International Database Engineering & Applications Symposium*. ACM, pp. 364–369.

**Dunietz J.**, **Levin L.** and **Carbonell J.** (2015). Annotating causal language using corpus lexicography of constructions. In *The 9th Linguistic Annotation Workshop held in Conjunction with NAACL 2015*, vol. 188.

**Dunietz J.**, **Levin L.** and **Carbonell J.** (2017). The BECauSE corpus 2.0: annotating causality and overlapping relations. In *LAW XI 2017*, vol. 95.

**Eisenschlos J.**, **Ruder S.**, **Czapla P.**, **Kadras M.**, **Gugger S.** and **Howard J.** (2019). Multifit: efficient multi-lingual language model fine-tuning. 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. 5706–5711.

**Fajcik M.**, **Jon J.**, **Docekal M.** and **Smrz P.** (2020). But-fit at semeval-2020 task 5: automatic detection of counterfactual statements with deep pre-trained language representation models. arXiv preprint arXiv:2007.14128.

**Girju R.** (2003). Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering - Volume 12*. Association for Computational Linguistics, pp. 76–83.

**Girju R.** and **Moldovan D.** (2002). Mining answers for causation questions. In *AAAI Symposium*.

**Gordon A.S.**, **Kozareva Z.** and **Roemmele M.** (2012). Semeval-2012 task 7: choice of plausible alternatives: an evaluation of commonsense causal reasoning. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval'12, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 394–398.

**Grivaz C.** (2012). *Automatic Extraction of Causal Knowledge from Natural Language Texts*. PhD Thesis, University of Geneva.

**Gui L.**, **Xu R.**, **Lu Q.**, **Wu D.** and **Zhou Y.** (2016). Emotion cause extraction, a challenging task with corpus construction. In *Chinese National Conference on Social Media Processing*. Springer, pp. 98–109.

**Gui L.**, **Yuan L.**, **Xu R.**, **Liu B.**, **Lu Q.** and **Zhou Y.** (2014). Emotion cause detection with linguistic construction in Chinese Weibo text. In *Natural Language Processing and Chinese Computing*. Springer, pp. 457–464.

**Hall N.** and **Paul L.** (2013). *Causation: A User's Guide*. Oxford: Oxford University Press.

**Hashimoto C.**, **Torisawa K.**, **Kloetzer J.** and **Oh J.-H.** (2015). Generating event causality hypotheses through semantic relations. In *AAAI*, pp. 2396–2403.

**Hashimoto C.**, **Torisawa K.**, **Kloetzer J.**, **Sano M.**, **Varga I.**, **Oh J.-H.** and **Kidawara Y.** (2014). Toward future scenario generation: extracting event causality exploiting semantic relation, context, and association features. In ACL (1), pp. 987–997.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

**Hassanzadeh O.**, **Bhattacharjya D.**, **Feblowitz M.**, **Srinivas K.**, **Perrone M.**, **Sohrabi S.** and **Katz M.** (2020). Causal knowledge extraction through large-scale text mining. In *AAAI*, pp. 13610–13611.

**Hendrickx I.**, **Kim S.N.**, **Kozareva Z.**, **Nakov P.**, **Ó Séaghdha D.**, **Padó S.**, **Pennacchiotti M.**, **Romano L.** and **Szpakowicz S.** (2009). Semeval-2010 task 8: multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, SEW'09, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 94–99.

**Hidey C.** and **McKeown K.** (2016). Identifying causal relations using parallel Wikipedia articles. In *Proceedings of the Association of Computational Linguistics.*

**Higashinaka R.** and **Isozaki H.** (2008a). Automatically acquiring causal expression patterns from relation-annotated corpora to improve question answering for why-questions. *ACM Transactions on Asian Language Information Processing (TALIP)* 7(2), 6.

**Higashinaka R.** and **Isozaki H.** (2008b). Corpus-based question answering for why-questions. In *IJCNLP*, pp. 418–425.

**Hitchcock C.R.** (1995). The mishap at Reichenbach fall: singular vs. general causation. *Philosophical Studies* **78**(3), 257–291.

**Hochreiter S.** and **Schmidhuber J.** (1997). Long short-term memory. *Neural Computation* **9**(8), 1735–1780.

**Hu Z.**, **Rahimtoroghi E.** and **Walker M.** (2017). Inference of fine-grained event causality from blogs and films. In *Proceedings of the Events and Stories in the News Workshop*, Vancouver, Canada. Association for Computational Linguistics, pp. 52–58.

**Ishii H.**, **Ma Q.** and **Yoshikawa M.** (2010a). Causal network construction to support understanding of news. In HICSS, pp. 1–10.

**Ishii H.**, **Ma Q.** and **Yoshikawa M.** (2010b). Causal network construction to support understanding of news. In *Proceedings of the Annual Hawaii International Conference on System Sciences.*

**Ittoo A.** and **Bouma G.** (2011a). Extracting explicit and implicit causal relations from sparse, domain-specific texts. In *International Conference on Application of Natural Language to Information Systems.* Springer, pp. 52–63.

**Ittoo A.** and **Bouma G.** (2011b). Extracting explicit and implicit causal relations from sparse, domain-specific texts. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*, vol. 6716, pp. 52–63.

**Izumi K.** and **Sakaji H.** (2019). Economic causal-chain search using text mining technology. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pp. 61–65.

**Jain L.C.** and **Medsker L.R.** (1999). *Recurrent Neural Networks: Design and Applications*, 1st Edn. Boca Raton, FL, USA: CRC Press, Inc.

**Jastrzebski S.**, **Lesniak D.** and **Czarnecki W.M.** (2017). How to evaluate word embeddings? on importance of data efficiency and simple supervised tasks. CoRR, abs/1702.02170.

**Jensen F.V.** (2001). Causal and Bayesian networks. In *Bayesian Networks and Decision Graphs*. Springer, pp. 3–34.

**Jin X.**, **Wang X.**, **Luo X.**, **Huang S.** and **Gu S.** (20200. Inter-sentence and implicit causality extraction from chinese corpus. In **Lauw H.W.**, **Wong R.C.-W.**, **Ntoulas A.**, **Lim E.-P.**, **Ng S.-K.** and **Pan S.J.** (eds), *Advances in Knowledge Discovery and Data Mining*. Cham: Springer International Publishing, pp. 739–751.

**Joskowicz L.**, **Ksiezyck T.** and **Grishman R.** (1989). Deep domain models for discourse analysis. In *AI Systems in Government Conference, 1989. Proceedings of the Annual*, pp. 195–200.

**Kaneko K.** and **Bekki D.** (2014a). Building a japanese corpus of temporal-causal-discourse structures based on sdrt for extracting causal relations. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pp. 33–39.

**Kaneko K.** and **Bekki D.** (2014b). Toward a discourse theory for annotating causal relations in japanese. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pp. 460–469.

**Kang D.**, **Gangal V.**, **Lu A.**, **Chen Z.** and **Hovy E.H.** (20170. Detecting and explaining causes from text for a time series event. In **Palmer M.**, **Hwa R.** and **Riedel S.** (eds), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9–11, 2017*. Association for Computational Linguistics, pp. 2758–2767.

**Khoo C.**, **Chan S.** and **Niu Y.** (2002). The many facets of the cause-effect relation. In **Green R.**, **Bean C.** and **Myaeng S.** (eds), *The Semantics of Relationships*. Information Science and Knowledge Management, vol. 3. Netherlands: Springer, pp. 51–70.

**Khoo C.S.**, **Myaeng S.H.** and **Oddy R.N.** (2001). Using cause-effect relations in text to improve information retrieval precision. *Information Processing & Management* **37**(1), 119–145.

**Khoo C.S.-G.** (1996). *Automatic Identification of Causal Relations in Text and their Use for Improving Precision in Information Retrieval*. PhD Thesis, Syracuse University.

**Kilicoglu H.** (2016). Inferring implicit causal relationships in biomedical literature. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pp. 46–55.

**Kim H.D.**, **Castellanos M.**, **Hsu M.**, **Zhai C.**, **Rietz T.** and **Diermeier D.** (2013). Mining causal topics in text data: iterative topic modeling with time series feedback. In *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management*, CIKM'13, New York, NY, USA. ACM, pp. 885–890.

**Kim H.D.**, **Zhai C.**, **Rietz T.A.**, **Diermeier D.**, **Hsu M.**, **Castellanos M.** and **Ceja Limon C.A.** (2012). Incatomi: integrative causal topic miner between textual and non-textual time series data. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM'12, New York, NY, USA. ACM, pp. 2689–2691.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

**Kolomiyets O.** and **Moens M.-F.** (2011). A survey on question answering technology from an information retrieval perspective. *Information Sciences* **181**(24), 5412–5434.

**Krishnan A.**, **Sligh J.**, **Tinsley E.**, **Crohn N.**, **Bandos J.**, **Bush H.**, **Depasquale J.** and **Palakal M.** (2014). Causal association mining from geriatric literature. In *2014 IEEE International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, pp. 226–230.

**Kumar S.** (2017). A survey of deep learning methods for relation extraction. arXiv preprint arXiv:1705.03645.

**Kunneman F.A.** and **van den Bosch A.** (2012). Leveraging unscheduled event prediction through mining scheduled event tweets. In *24th Benelux Conference on Artificial Intelligence*. Maastricht:[sn].

**Kyriakakis M.**, **Androutsopoulos I.**, **Saudabayev A.** and **Ginés i Ametllé J.** (2019). Transfer learning for causal sentence detection. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, Florence, Italy. Association for Computational Linguistics, pp. 292–297.

**Lee S.Y.M.**, **Chen Y.** and **Huang C.-R.** (2010). A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Association for Computational Linguistics, pp. 45–53.

**Lee S.Y.M.**, **Chen Y.**, **Huang C.-R.** and **Li S.** (2013). Detecting emotion causes with a linguistic rule-based approach. *Computational Intelligence* **29**(3), 390–416.

**Levin B.** (1986). Causation: The perspective from resultatives. In *The New York Times*, vol. **8**.

**Levin B.** (1993). *English Verb Classes and Alternations*. University of Chicago Press.

**Li W.**, **Wu M.**, **Lu Q.**, **Xu W.** and **Yuan C.** (2006). Extractive summarization using inter-and intra-event relevance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 369–376.

**Li W.** and **Xu H.** (2014). Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications* 41(4), 1742–1749.

**Li X.**, **Xie H.**, **Chen L.**, **Wang J.** and **Deng X.** (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems* **69**, 14–23.

**Li Z.**, **Ding X.** and **Liu T.** (2018). Constructing narrative event evolutionary graph for script event prediction. In *Proceedings of the IJCAI 2018*.

**Li Z.**, **Li Q.**, **Zou X.** and **Ren J.** (2019). Causality extraction based on self-attentive BiLSTM-CRF with transferred embeddings. arXiv preprint arXiv:1904.07629.

**Li Z.**, **Li Q.**, **Zou X.** and **Ren J.** (2021). Causality extraction based on self-attentive biLSTM-CRF with transferred embeddings. *Neurocomputing* **423**, 207–219.

**Liu J.**, **Chen Y.** and **Zhao J.** (2020). Knowledge enhanced event causality identification with mention masking generalizations. In **Bessiere**, **C.** (ed), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. International Joint Conferences on Artificial Intelligence Organization, pp. 3608–3614.

**Liu X.**, **He P.**, **Chen W.** and **Gao J.** (2019). Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4487–4496.

**Lorenz G.** (1999). Learning to cohere: causal links in native vs. non-native argumentative writing. In *Pragmatics and Beyond New Series*, pp. 55–76.

**Luo Z.**, **Sha Y.**, **Zhu K.Q.**, **Hwang S.-w.** and **Wang Z.** (2016). Commonsense causal reasoning between short texts. In *KR*, pp. 421–431.

**Mackie, J.L.** (1965). Causes and conditions. *American Philosophical Journal* **2**, 245–255.

**Mackie J.L.** (1974). *The Cement of the Universe: A Study of Causation*. Oxford: Clarendon Press.

**Maekawa K.**, **Yamazaki M.**, **Ogiso T.**, **Maruyama T.**, **Ogura H.**, **Kashino W.**, **Koiso H.**, **Yamaguchi M.**, **Tanaka M.** and **Den Y.** (2014). Balanced corpus of contemporary written japanese. *Language Resources and Evaluation* **48**(2), 345–371.

**Marcu D.** and **Echihabi A.** (2002). An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL'02, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 368–375.

**Mariko D.**, **Akl H.A.**, **Labidurie E.**, **Durfort S.**, **De Mazancourt H.** and **El-Haj M.** (2020). Financial document causality detection shared task (fincausal 2020). arXiv preprint arXiv:2012.02505.

**Mehrabi S.**, **Krishnan A.**, **Tinsley E.**, **Sligh J.**, **Crohn N.**, **Bush H.**, **Depasquale J.**, **Bandos J.** and **Palakal M.** (2013). Event causality identification using conditional random field in geriatric care domain. In *Proceedings - 2013 12th International Conference on Machine Learning and Applications, ICMLA 2013*, vol. 1, pp. 339–343.

**Mellor D.** (1998). *The Facts of Causation. International Library of Philosophy*, Psychology, and Scientific Method. Routledge.

**Mihaila C.** and **Ananiadou S.** (2013). What causes a causal relation? detecting causal triggers in biomedical scientific discourse. In ACL (Student Research Workshop), pp. 38–45.

**Mihăilă C.** and **Ananiadou S.** (2014). Semi-supervised learning of causal relations in biomedical scientific discourse. *Biomedical Engineering Online* 13(Suppl 2), S1.

**Mikolov T.**, **Grave É.**, **Bojanowski P.**, **Puhrsch C.** and **Joulin A.** (2018). Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

**Mikolov T.**, **Sutskever I.**, **Chen K.**, **Corrado G.S.** and **Dean, J.** (2013). Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems, pp. 3111–3119.

**Miller G.A.** (1995). WordNet: a lexical database for English. *Communications of the ACM* **38**, 39–41.

**Mirza P.** (2016). Extracting temporal and causal relations between events. arXiv preprint arXiv:1604.08120.

**Mirza P.**, **Sprugnoli R.**, **Tonelli S.** and **Speranza M.** (2014). Annotating causality in the tempeval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pp. 10–19.

**Mirza P.** and **Tonelli S.** (2014). An analysis of causality between events and its relation to temporal information. In COLING, pp. 2097–2106.

**Mirza P.** and **Tonelli S.** (20160. Catena: causal and temporal relation extraction from natural language texts. In *The 26th International Conference on Computational Linguistics*. ACL, pp. 64–75.

**Mostafazadeh N.**, **Grealish A.**, **Chambers N.**, **Allen J.** and **Vanderwende L.** (2016). Caters: causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the 4th Workshop on Events: Definition, Detection, Coreference, and Representation*, pp. 51–61.

**Mou L.**, **Meng Z.**, **Yan R.**, **Li G.**, **Xu Y.**, **Zhang L.** and **Jin Z.** (2016). How transferable are neural networks in NLP applications? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 479–489.

**Mulkar-Mehta R.**, **Gordon A.S.**, **Hovy E.** and **Hobbs J.R.** (2011a). Causal markers across domains and genres of discourse. In *The 6th International Conference on Knowledge Capture*, Banff, Alberta, Canada.

**Mulkar-Mehta R.**, **Welty C.**, **Hobbs J.R.** and **Hovy E.** (2011b). Using granularity concepts for discovering causal relations. In *Proceedings of the FLAIRS Conference*.

**Mulkar-Mehta R.**, **Welty C.A.**, **Hobbs J.R.** and **Hovy**, **E.H.** (2011c). Using part-of relations for discovering causality. In *FLAIRS Conference*.

**Murphy K.P.** (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

**Neeleman A.**, and **Van de Koot H.** (2012). The linguistic expression of causation. In *The Theta System: Argument Structure at the Interface*, vol. 20.

**Ning Q.**, **Feng Z.**, **Wu H.** and **Roth D.** (2018). Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, pp. 2278–2288.

**Nwaike K.** and **Jiao L.** (2020). Counterfactual detection meets transfer learning. arXiv preprint arXiv:2005.13125.

**O'Gorman T.**, **Wright-Bettner K.** and **Palmer M.** (2016). Richer event description: integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pp. 47–56.

**Oh J.-H.**, **Torisawa K.**, **Hashimoto C.**, **Sano M.**, **De Saeger S.** and **Ohtake**, **K.** (2013). Why-question answering using intra- and inter-sentential causal relations. In *ACL (1)*, pp. 1733–1743.

**Ojha A.A.**, **Garg R.**, **Gupta S.** and **Modi A.** (2020). Iitk-rsa at semeval-2020 task 5: detecting counterfactuals. arXiv e-prints, pp. arXiv–2007.

**Onyshkevych B.** (1993). Template design for information extraction. In *Proceedings of the 5th Conference on Message Understanding*. Association for Computational Linguistics, pp. 19–23.

**Ovchinnikova E.**, **Montazeri N.**, **Alexandrov T.**, **Hobbs J.R.**, **McCord**, **M.C.** and **Mulkar-Mehta**, **R.** (2014). Abductive reasoning with a large knowledge base for discourse processing. In *Computing Meaning*. Springer, pp. 107–127.

**Ovchinnikova E.**, **Vieu L.**, **Oltramari A.**, **Borgo S.** and **Alexandrov T.** (2010). Data-driven and ontological analysis of framenet for natural language reasoning. In *LREC*.

**Page L.**, **Brin S.**, **Motwani R.** and **Winograd T.** (1999). The pagerank citation ranking: bringing order to the web. Technical report, Stanford InfoLab.

**Palmer M.**, **Bonial C.** and **Hwang J.D.** (2017). Verbnet: capturing english verb behavior, meaning and usage. In The Oxford Handbook of Cognitive Science, pp. 315–336.

**Papanikolaou Y.**, **Roberts I.** and **Pierleoni A.** (2019). Deep bidirectional transformers for relation extraction without supervision. In *EMNLP-IJCNLP 2019*, vol. 67.

**Pearl J.** (2012). The do-calculus revisited. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 3–11.

**Pearl J.** and **Mackenzie D.** (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books.

**Pechsiri C.** and **Kawtrakul A.** (2007). Mining causality from texts for question answering system. *IEICE Transactions on Information and Systems* **90**(10), 1523–1533.

**Pennington J.**, **Socher R.** and **Manning C.D.** (2014). Glove: global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

**Peters M.**, **Neumann M.**, **Iyyer M.**, **Gardner M.**, **Clark C.**, **Lee K.** and **Zettlemoyer L.** (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana. Association for Computational Linguistics, pp. 2227–2237.

**Pichotta K.** and **Mooney R.** (2016). Using sentence-level LSTM language models for script inference. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 279–289.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

**Ponti E.M.** and **Korhonen A.** (2017). Event-related features in feedforward neural networks contribute to identifying causal relations in discourse. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pp. 25–30.

**Preethi P.G.**, **Uma V.** and **Ajit K.** (2015). Temporal sentiment analysis and causal rules extraction from tweets for event prediction. *Procedia Computer Science*, **48**, 84–89.

**Psillos S.** (2007). Causal explanation and manipulation. In *Rethinking Explanation*. Springer, pp. 93–107.

**Puente C.**, **Garrido E.** and **Olivas J.A.** (2013a). Answering questions by means of causal sentences. In *International Conference on Flexible Query Answering Systems*. Springer, pp. 91–99.

**Puente C.**, **Olivas J.** and **Prado I.** (2014). Summarizing information by means of causal sentences. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, vol. 1. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

**Puente C.**, **Olivas J.A.**, **Garrido E.** and **Seisdedos R.** (2013b). Creating a natural language summary from a compressed causal graph. In *IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), 2013 Joint*. IEEE, pp. 513–518.

**Puente C.**, **Sobrino A.**, **Olivas J.** and **Garrido E.** (2016). Summarizing information by means of causal sentences through causal graphs. *Journal of Applied Logic* **24**, 3–14.

**Puente C.**, **Villa-Monte A.**, **Lanzarini L.**, **Sobrino A.** and **Olivas J.A.** (2017). Evaluation of causal sentences in automated summaries. In 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE, pp. 1–6.

**Qiu J.**, **Xu L.**, **Zhai J.** and **Luo L.** (2017). Extracting causal relations from emergency cases based on conditional random fields. *Procedia Computer Science* **112**, 1623–1632.

**Radford A.**, **Wu J.**, **Child R.**, **Luan D.**, **Amodei D.** and **Sutskever I.** (2019). Language models are unsupervised multitask learners. Available at https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (accessed October 2020).

**Radinsky K.**, **Davidovich S.** and **Markovitch S.** (2012). Learning causality for news events prediction. In *Proceedings of the 21st International Conference on World Wide Web*. ACM, pp. 909–918.

**Radinsky K.** and **Horvitz E.** (2013). Mining the web to predict future events. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM'13. ACM, pp. 255–264.

**Rashmi P.**, **Nihkil D.**, **Alan L.**, **Eleni M.**, **Robaldo L.**, **Aravind J.**, **Bonnie W.** (2008). The penn discourse treebank 2.0. In *Lexical Resources and Evaluation Conference.*

**Riaz M.** (2010). *An Unsupervised Approach to Identifying Causal Relations from Relevant Scenarios*. Master's Thesis, University of Illinois.

**Riaz M.** and **Girju R.** (2010). Another look at causality: discovering scenario-specific contingency relationships with no supervision. In *2010 IEEE Fourth International Conference on Semantic Computing (ICSC)*. IEEE, pp. 361–368.

**Riaz M.** and **Girju R.** (2014). Recognizing causality in verb-noun pairs via noun and verb semantics. In *Proceedings of the Workshop on Computational Approaches to Causality in Language EACL 2014*. The Association for Computer Linguistics.

**Rotmensch M.**, **Halpern Y.**, **Tlimat A.**, **Horng S.** and **Sontag D.** (2017). Learning a health knowledge graph from electronic medical records. *Scientific Reports* **7**(1), 1–11.

**Russell S.J.**, **Norvig P.** and **Norvig P.** (2003). *Artificial Intelligence: Prentice Hall Series in Artificial Intelligence*. Upper Saddle River, NJ: Pearson Education.

**Sadek J.** (2013). Automatic detection of arabic causal relations. In **Metais E.**, **Meziane F.**, **Saraee M.**, **Sugumaran V.** and **Vadera S.** (eds), *Natural Language Processing and Information Systems*, Lecture Notes in Computer Science, vol. 7934. Berlin, Heidelberg: Springer, pp. 400–403.

**Sakai H.** and **Masuyama S.** (2007). Extraction of cause information from newspaper articles concerning business performance. In Artificial Intelligence and Innovations 2007: from Theory to Applications, pp. 205–212.

**Sakaji H.**, **Sakai H.** and **Masuyama S.** (2008a). *Automatic Extraction of Basis Expressions That Indicate Economic Trends.* Berlin, Heidelberg: Springer, pp. 977–984.

**Sakaji H.**, **Sekine S.** and **Masuyama S.** (2008b). Extracting causal knowledge using clue phrases and syntactic patterns. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*, LNAI, vol. 5345, pp. 111–122. cited By (since 1996)1.

**Sanchez-Graillet O.** and **Poesio M.** (2004). Acquiring Bayesian networks from text. In *LREC*.

**Sastre J.**, **Zaman F.**, **Duggan N.**, **McDonagh C.** and **Walsh P.** (2020). A deep learning knowledge graph approach to drug labelling. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, pp. 2513–2521.

**Schuler K.K.** (2005). *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. PhD Thesis, University of Pennsylvania, Philadelphia, PA, USA. AAI3179808.

**Shapiro S.C.** (1992). Semantic networks. *In Encyclopedia of Artificial Intelligence*, 2nd Edn. New York, NY, USA: John Wiley & Sons, Inc.

**Sharp R.**, **Surdeanu M.**, **Jansen P.**, **Clark P.** and **Hammond M.** (2016). Creating causal embeddings for question answering with minimal supervision. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas. Association for Computational Linguistics, pp. 138–148.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

**Sizov G.** and **Öztürk P.** (2013). Automatic extraction of reasoning chains from textual reports. In *Proceedings of TextGraphs-8 Graph-based Methods for Natural Language Processing*, pp. 61–69.

**Sogaard A.** (2013). *Semi-Supervised Learning and Domain Adaptation in Natural Language Processing*, 1st Edn. Morgan & Claypool Publishers.

**Son Y.**, **Bayas N.** and **Schwartz H.A.** (2018). Causal explanation analysis on social media. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3350–3359.

**Son Y.**, **Buffone A.**, **Raso J.**, **Larche A.**, **Janocko A.**, **Zembroski K.**, **Schwartz H.A.** and **Ungar L.** (2017). Recognizing counterfactual thinking in social media texts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 654–658.

**Tirunagari S.**, **Hanninen M.**, **Stanhlberg K.** and **Kujala P.** (2012). Mining causal relations and concepts in maritime accidents investigation reports. *International Journal of Innovative Research and Development* **1**(10), 548–566.

**UzZaman N.**, **Llorens H.**, **Derczynski L.**, **Allen J.**, **Verhagen M.** and **Pustejovsky J.** (2013). SemEval-2013 task 1: TempEval-3: evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (∗SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, USA. Association for Computational Linguistics, pp. 1–9.

**VanVactor, J.D.** (2010). Health care logistics response in a disaster. *Journal of Homeland Security and Emergency Management* **7**(1), 1–17.

**Vaswani A.**, **Shazeer N.**, **Parmar N.**, **Uszkoreit J.**, **Jones L.**, **Gomez A.N.**, **Kaiser Ł.** and **Polosukhin I.** (2017). Attention is all you need. In Advances in Neural Information Processing Systems, pp. 5998–6008.

**Vendler Z.** (1967). Causal relations. *The Journal of Philosophy* **64**(21), 704–713.

**Weber N.**, **Rudinger R.** and **Van Durme B.** (2020). Causal inference of script knowledge. arXiv, arXiv–2004.

**Wolf T.**, **Debut L.**, **Sanh V.**, **Chaumond J.**, **Delangue C.**, **Moi A.**, **Cistac P.**, **Rault T.**, **Louf R.**, **Funtowicz M.**, et al. (2019). Transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.

**Woodward J.** (2005). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

**Wu J.-L.**, **Yu L.-C.** and **Chang P.-C.** (2012). Detecting causality from online psychiatric texts using inter-sentential language patterns. *BMC Medical Informatics and Decision Making* **12**(1), 1–10.

**Xu R.**, **Hu J.**, **Lu Q.**, **Wu D.** and **Gui L.** (2017). An ensemble approach for emotion cause detection with event extraction and multi-kernel svms. *Tsinghua Science and Technology* **22**(6), 646–659.

**Yang B.**, **Wu J.** and **Hattori G.** (2020a). A transfer learning method of data collection for dialogue response generation concerning causal relation. In *Proceedings of the Conference of the Japanese Society for Artificial Intelligence*. Springer.

**Yang X.**, **Obadinma S.**, **Zhao H.**, **Zhang Q.**, **Matwin S.** and **Zhu X.** (2020b). Semeval-2020 task 5: counterfactual recognition. arXiv preprint arXiv:2008.00563.

**Yang Y.**, **Wei Z.**, **Chen Q.** and **Wu L.** (2019a). Using external knowledge for financial event prediction based on graph neural networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 2161–2164.

**Yang Z.**, **Dai Z.**, **Yang Y.**, **Carbonell J.**, **Salakhutdinov R.R.**, and **Le Q.V.** (2019b). Xlnet: generalized autoregressive pretraining for language understanding. In Advances in Neural Information Processing Systems, pp. 5753–5763.

**Yu B.**, **Li Y.** and **Wang J.** (2019). Detecting causal language use in science findings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4656–4666.

**Yu B.**, **Wang J.**, **Guo L.** and **Li Y.** (2020). Measuring correlation-to-causation exaggeration in press releases. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online). International Committee on Computational Linguistics, pp. 4860–4872.

**Yu H.** (2020a). Health causal probability knowledge graph: another intelligent health knowledge discovery approach. In *2020 7th International Conference on Bioinformatics Research and Applications*, pp. 49–58.

**Yu H.Q.** (2020b). Dynamic causality knowledge graph generation for supporting the chatbot healthcare system. In *Proceedings of the Future Technologies Conference*. Springer, pp. 30–45.

**Zhang Y.**, **Jatowt A.** and **Tanaka K.** (2016a). Causal relationship detection in archival collections of product reviews for understanding technology evolution. *ACM Transactions on Information Systems (TOIS)* **35**(1), 3.

**Zhang Y.**, **Jatowt A.** and **Tanaka K.** (2016b). Detecting evolution of concepts based on cause-effect relationships in online reviews. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pp. 649–660.

**Zhao S.**, **Wang Q.**, **Massung S.**, **Qin B.**, **Liu T.**, **Wang B.** and **Zhai C.** (2017). Constructing and embedding abstract event causality networks from text snippets. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, pp. 335–344.

**Zhou P.**, **Shi W.**, **Tian J.**, **Qi Z.**, **Li B.**, **Hao H.** and **Xu B.** (2016). Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 207–212.

https://doi.org/10.1017/S135132492100036X Published online by Cambridge University Press

**Zuo X.**, **Chen Y.**, **Liu K.** and **Zhao J.** (2020a). Knowdis: knowledge enhanced data augmentation for event causality detection via distant supervision. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 1544–1550.

**Zuo X.**, **Chen Y.**, **Liu K.** and **Zhao J.** (2020b). Towards causal explanation detection with pyramid salient-aware network. In *China National Conference on Chinese Computational Linguistics*. Springer, pp. 113–128.

**Cite this article:** Drury B, Gonçalo Oliveira H and de Andrade Lopes A (2022). A survey of the extraction and applications of causal relations. *Natural Language Engineering* **28**, 361–400. https://doi.org/10.1017/S135132492100036X