Knowledge and Information Systems (2022) 64:1161–1186
https://doi.org/10.1007/s10115-022-01665-w

SURVEY PAPER

A survey on extraction of causal relations from natural language text

Jie Yang[1] · Soyeon Caren Han[1] · Josiah Poon[1]

Received: 23 January 2021 / Revised: 7 February 2022 / Accepted: 12 February 2022 /
Published online: 12 March 2022

Abstract
As an essential component of human cognition, cause–effect relations appear frequently in text, and curating cause–effect relations from text helps in building causal networks for predictive tasks. Existing causality extraction techniques include knowledge-based, statistical machine learning (ML)-based, and deep learning-based approaches. Each method has its advantages and weaknesses. For example, knowledge-based methods are understandable but require extensive manual domain knowledge and have poor cross-domain applicability. Statistical machine learning methods are more automated because of natural language processing (NLP) toolkits. However, feature engineering is labor-intensive, and toolkits may lead to error propagation. In the past few years, deep learning techniques attract substantial attention from NLP researchers because of its powerful representation learning ability and the rapid increase in computational resources. Their limitations include high computational costs and a lack of adequate annotated training data. In this paper, we conduct a comprehensive survey of causality extraction. We initially introduce primary forms existing in the causality extraction: explicit intra-sentential causality, implicit causality, and inter-sentential causality. Next, we list benchmark datasets and modeling assessment methods for causal relation extraction. Then, we present a structured overview of the three techniques with their representative systems. Lastly, we highlight existing open challenges with their potential directions.

Keywords Causality extraction · Explicit intra-sentential causality · Implicit causality · Inter-sentential causality · Deep learning

✉ Soyeon Caren Han
caren.han@sydney.edu.au

Jie Yang
jyan4704@uni.sydney.edu.au

Josiah Poon
josiah.poon@sydney.edu.au

[1] School of Computer Science, The University of Sydney, 1 Cleveland Street, Sydney, NSW 2006, Australia

1 Introduction

With the rapid growth of unstructured texts online, information extraction (IE) plays a vital role in NLP research. It automatically transforms and stores unstructured texts into machine readable data [20]. The complex syntax and semantics of natural language and its extensive vocabulary make IE a challenging task. IE is an aggregation of tasks, which includes named entity recognition (NER), relation extraction (RE), and event extraction. RE refers to extracted and classified semantic relationships, such as whole–part, product–producer, and cause–effect from text. Specifically, the cause–effect relation, which refers to a relationship between two entities *e1* and *e2* that the occurrence of *e1* results in the occurrence of *e2*, is essential in many areas. For example, in medicine, the decision to provide a treatment is based on the relationship that the treatment leads to an improvement in patient's condition. Or, the critical issues of whether a disease is the reason for a symptom depend on if there are cause–effect relation between them. Extracting such kinds of causal relations from the medical literature can support constructing a knowledge graph, which can assist doctors in quickly finding causality, like *diseases-cause-symptoms, diseases-bring-complications, treatments-improve-conditions*, and finally customize treatment plans. Similarly, extracting cause–effect relations from text, which is the study of causality extraction (CE), has received ongoing attention in media [3,13,19,46,73], biomedical [12,47,63,77], emergency management [78], etc.

The task of CE focuses on developing systems for identifying cause–effect relations between pairs of labeled nouns from text [5]. From the aspect of techniques, as shown in Fig. 1, there has been a considerable body of CE systems that can be divided into three groups: knowledge-based approaches, statistical ML-based approaches, and deep learning-based approaches. Alternatively, CE studies can be classified in terms of different representation patterns: explicit or implicit causality, intra- or inter-sentential causality. Explicit causality has relations that are connected by the following explicit causal connectives: (a) causal links (e.g., *so, hence, therefore, because of, on account of, because, as, since, the result was*); (b) causative verbs (e.g., *break, kill*); (c) resultative phrases; (d) conditional, i.e., *if...then...*; and (e) causative adverbs and adjectives [45]. Implicit causality means explicit causal valence is replaced by ambiguous connectives, e.g., *as, after* in the first four examples, or even without any connectives, as the last example in Table 1. Readers need to use background knowledge to analyzing and reasoning if there is causality in the text. In intra-sentential causality, the "cause" and the "effect" lie in a single sentence, while in inter-sentential causality, the "cause" and the "effect" lie in different sentences. Most CE approaches, like [25,26,47,51,63,79], identify causality in the basic levels, which are explicit and/or intra-sentential forms.

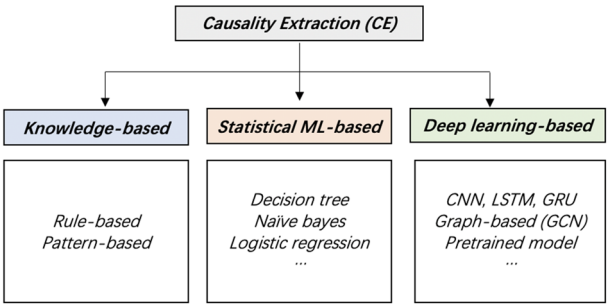

Fig. 1 Taxonomy of techniques

Table 1 Examples for implicit causality

Connectives	Sentences	Labels
As	There was no debate *as* the Senate passed the bill on to the House [10]	Causal
As	It has a fixed time, *as* collectors well know [10]	Non-causal
After	Bischoff in a round table discussion claimed he fired Austin *after* he refused to do a taping in Atlanta [61]	Causal
After	In stark contrast to his predecessor, five days *after* his election he spoke of his determination to do what he could to bring peace [61]	Non-causal
–	He derives great joy and happiness from cycling [5]	Causal

Table 2 The forms of causal relations

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

However, causality in many texts is implicit and/or inter-sentential conditions, which are more complicated than basic kinds of causality. Table 2 lists three examples, which include the sentences, causality forms, and the causality pairs.

The rest of the article is structured as follows. We review in detail of previous surveys in Sect. 2. The benchmark datasets and evaluation metrics for CE system are presented in Sects. 3 and 4, respectively. Then, we survey representative CE systems and summarize them in Sects. 5 and 6. We propose three open problems of the CE task with their potential solutions in Sect. 7, and the conclusion of this paper is in Sect. 8.

2 Previous surveys

With limited exceptions, there is a notable paucity of surveys focusing specifically on CE. It may be because cause–effect is a common relation that researchers scale up to RE literature reviews. Examples include the generalized survey [96], detailed analyses of RE in the biomedical domain [41,101], and a survey about the application of distant supervision on RE [82]. From our point of view, however, CE is different from RE, as the former task is a binary classification while the later is multiple classification problem. Meanwhile, the two tasks focus on different kinds of linguistic patterns or features. For example, the punctuation feature can be used in RE to indicate the relation of *Description* and *Attribution*, but it is useless in the task of CE [44]. Also, RE faces the challenge of extracting relations on open-domain corpora, that is, the relation types may not be pre-defined [90], while the target of CE is clear and there are no new relation types.

In 2016, Asghar [2] separates CE applications into non-statistical techniques, and statistical and machine learning techniques. Besides reviewing previous approaches, another contribution is the analysis of strengths and weaknesses of the two categories. Early non-statistical methods suffer from constructing annotated linguistic and syntactic patterns manually, while ML-based systems can utilize a small set of seed patterns with algorithms to find these language patterns automatically. Also, most non-statistical models restricted their corpora to a particular domain with a specific text type (e.g., narrative, prose, drama). In comparison, the statistical ML techniques provide better generalization to other domains and types of text. Meanwhile, unlike non-statistical architectures that only extracted explicit cause–effect relations, a large number of ML systems (e.g., [9,80,83,92]) have the capability to explore implicit relations. In the same year with the study of [2], Barik et al. [4] categorize existing CE approaches into four groups: using handcrafted patterns, using semiautomatic causal patterns, using supervised learning, and statistical methods. From their point of view, instead of using manually linguistic clues and domain knowledge, semiautomatic learning acquires lexico-syntactic patterns from a larger corpus automatically. Then, these patterns are used to identify in-domain causal relations or evaluate causal patterns in a semiautomated way. For the supervised learning, there are a large number of corpora that required labeled prior to modeling. The above two surveys provide comprehensive reviews of CE, one of their limitations is the lack of review about recent developments in the field, especially deep learning. Luckily, we will review both of the traditional and modern methods in Sect. 5.

3 Benchmark datasets

As we all know that data is the foundation of experiment. There is a number of datasets which have been previously used for evaluating CE models. In this section, we describe four datasets from general domain and two datasets from biomedical domain and summarize them in terms of their causality sizes, sources, available condition, balanced condition (X and - represent balanced and imbalanced, respectively), and related works in Table 3.

- **SemEval-2007 task 4** It is part of SemEval (Semantic Evaluation), the 4th edition of the semantic evaluation event [27]. This task provides a dataset for classifying semantic relations between two nominals. Within the set of seven relations, the organizers split the *Cause–Effect* examples into 140 training with 52.0% positive data, and 80 test with 51.0% positive data. This dataset has the following advantages: (a) Strong reputation. SemEval is one of the most influential, largest scale natural language semantic evaluation competition. As of 2020, SemEval has been successfully held for fourteen sessions and has a high impact in both industry and academia. (b) Easily accessible. Each relation example with the annotated results is collected in a separate TXT file, which can also reduce the workload of data preprocessing. On the contrary, the main limitation is the small data amount that 140 training and 80 test examples are far from meeting the needs for developing a CE system.
- **SemEval-2010 task 8** Unlike its predecessor, SemEval-2007 Task 4, which has an independent binary-labeled dataset for each kind of relation, this is a multi-classification task in which relation label for each sample is one of nine kinds of relations [34]. Within the 10,717 annotated examples, there are 1003 training with 13.0% positive data, and 328 test with 12.0% positive data. This small sample amount and imbalanced condition are the major limitations of this dataset.

Table 3 Benchmark datasets

Datasets	Published years	Causality sizes	Sources	Availability	Balanced	Related works
SemEval-2007 task 4	2007	220	Wikipedia	Publicly available[a]	X	[5,28]
SemEval-2010 Task 8	2010	1331	Wikipedia	Publicly available[b]	–	[51,56,69,83, 89,91,98,99]
PDTB 2.0	2018	9190	WSJ	License required[c]	–	[15,35,52,57, 61,73,81]
TACRED	2018	269	Newswire, Web	License required[d]	–	[97,98]
BioInfer	2007	1461	PubMed	Publicly available[e]	X	[1,14]
ADE	2012	6821	PubMed	Publicly available[f]	–	[6,7,33,42,55, 88,100]

[a] https://sites.google.com/site/semeval2007task4/data
[b] https://github.com/sahitya0000/Relation-Classification/tree/master/corpus
[c] https://catalog.ldc.upenn.edu/LDC2008T05
[d] https://catalog.ldc.upenn.edu/LDC2018T24
[e] http://mars.cs.utu.fi/BioInfer/?q=download
[f] https://sites.google.com/site/adecorpus/

– **PDTB 2.0** The second release of the penn discourse treebank (PDTB) dataset from Prasad et al. [74] is the largest annotated corpus of discourse relations. It includes 72,135 non-causal and 9190 causal examples from 2312 Wall Street Journal (WSJ) articles. In addition, there is a type of implicit relation in the dataset known as AltLex (Alternative lexicalization) corpus, in which causal meanings are not expressed by explicit causal lexical markers. However, the authors store PDTB in a complex way that researchers need to use tools to convert it into easy-to-operate files.
– **TACRED** Similar to SemEval, the Text Analysis Conference (TAC) is a series of evaluation workshops about NLP research. The TAC Relation Extraction Dataset (TACRED) contains 106,264 newswire and online text that have been collected from the TAC KBP challenge.[1] during the year from 2009 to 2014 [97]. The sentences are annotated with person- and organization-oriented related type (e.g., *per:title, org:founded*). The main limitation of TACRED is the small number of examples that there are only 269 *cause_of_death* instances available for CE task.

The above four corpora are collected from large general-purpose texts, like English Wikipedia and WSJ. At the same time, datasets in specific domains are needed to train and evaluate specific CE systems. Here, we list two causality datasets in the biomedical domain.

– **BioInfer** Pyysalo et al. [75] introduce an annotated corpus, BioInfer (Bio Information Extraction Resource), which contains 1100 sentences with the relations of genes, proteins, and RNA from biomedical publications. There are 2662 relations in the 1100 sentences, of these 1461 (54.9%) are causal-effect. The original data is collected in detail in the XML form, including sentence with entity markup.
– **ADE** The corresponding ADE task aims to extract two entities (drugs and diseases) and relations about drugs with their adverse effects (ADEs) [33,55]. Dataset in the task is

[1] https://www.ldc.upenn.edu/collaborations/current-projects/tac-kbp.

collected from 1644 PubMed abstracts, in which 6821 sentences have at least one ADE relation, and 16,695 sentences are annotated as non-ADE sentences. Annotators only label drugs and diseases in the ADE sentences, so some studies, like [55], only use the 6821 sentences in the experiments.

4 Evaluation metrics

To evaluate the performance of a CE system, the following four metrics are commonly used:

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{2}$$

$$\text{F-score} = \frac{2 * \text{TP}}{(2 * \text{TP} + \text{FN} + \text{FP})} \tag{3}$$

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})} \tag{4}$$

As many researchers define their CE systems as relation extraction tasks, that is, to determine whether the annotated causal pair in the input text has causality. Within their evaluation metrics, TP (true positive) is the number of correctly identified causal pairs. FP (false positive) refers to the number of causal pairs identified as non-causal pairs. TN (true negative) is the number of correctly identified non-causal pairs, and FN (false negative) is the number of non-causal pairs that are identified as causal pairs.

Accuracy and F-score have been (and still are) among the most popular adopted metrics in most classification tasks. However, they may generate overoptimistic, misleading results on imbalanced datasets, as they failed to consider the ratio between positive and negative classes [16]. In contrast, Matthews correlation coefficient (MCC) [62] views two classes are equal importance. It is high only when the classifier is doing well in both positive and negative classes:

$$\text{MCC} = \frac{(\text{TP} * \text{TN} - \text{FP} * \text{FN})}{\sqrt{(\text{TP} + \text{FP}) * (\text{TP} + \text{FN}) * (\text{TN} + \text{FP}) * (\text{TN} + \text{FN})}} \tag{5}$$

The MCC has been used for classifier evaluation over imbalanced datasets, as the publication of [23,54,58].

The geometric mean [50], G-mean, also indicates the balance between performances on both classes. A poor performance in positive examples prediction will lead to a low G-mean value, even if negative instances are correctly classified by the classifier [31]:

$$\text{G-mean} = \sqrt{\frac{\text{TP}}{(\text{TP} + \text{FN})} * \frac{\text{TN}}{(\text{TN} + \text{FP})}} \tag{6}$$

The effectiveness of G-mean for classifier assessment over imbalanced datasets has been shown in many studies, like [24,43,84].

The entity labeling metrics is also applied to evaluate the models. For example, Khoo et al. [46] use average precision and recall to judge whether the model can identify both the boundary of cause and effect. Dasgupta et al. [21] compare F-score of labeling "C" (cause), "E" (effect), "CC" (causal connectives), and "N" (None) tags with baseline models. Compared with the relation extraction method, evaluate in a labeling way is more suitable

for these systems: Both cause and effect have more than one word, and there is no entity mask in the original sentence.

Meanwhile, some approaches evaluate their models based on their topics. The studies of [66–68] aim to recognize causality for finding proper answers in why-QA (question-answering) system. So the authors evaluate their models by precision of the top answer (P@N) and mean average precision (MAP), where P@N measures the number of questions that have correct answers in the top-N passages, and MAP measures the quality of the answer passages ranked by systems. Kim et al. [48] report causality confidence and topic purity to measure the quality for mining causality topics. For causality confidence, they use the p-value of the Granger causality testing [30] between two variables. For topic purity, they calculate the entropy of cause word distributions and normalize it to the [0, 100] range.

5 Causal relation extraction methods

Many researchers have devoted themselves to the study of causality extraction. In the following subsections, we summarize and classify existing methods for causality extraction, based on the underlying techniques and on the causality forms identified.

5.1 Knowledge-based approaches

Knowledge-based CE systems can be divided into pattern-based approaches and rule-based approaches. Some of the former systems express linguistic patterns by means of pre-defined graphical patterns, or keywords (e.g., *thanks to, because, lead to*). On the other hand, patterns can also be explored through sentence structure analyses, like lexico-semantic or syntactic analysis. These structure analysis techniques lead to performance improvements and additionally are more likely to extract implicit causality relations. As to rule-based approaches, while some systems (similarly to pattern-based approaches) rely on patterns or templates to identify candidate causal instances directly, other systems employ a set of procedures or heuristic algorithms on the syntactic structure of sentences.

In the following paragraphs we will review how the existing systems, described in the literature, employ knowledge-based techniques to extract causality in different forms.

5.1.1 Explicit intra-sentential causality

Garcia et al. [25] and Khoo et al. [47] use patterns to identify explicitly expressed causal relations, within a single sentence. The tool developed by Garcia and colleagues, COATIS, extracts causality from French texts through lexico-syntactic patterns based on 23 explicit causal verbs, like *provoke, disturb, result, lead to*. Due to the special attention given to the syntactic positions of causal verbs and their surrounding noun phrases, COATIS achieves a reasonable precision of 85.2%. Even though it can only be applied to small fragments of French text, it illustrates how to implement a domain-independent CE system via pre-defined patterns. Khoo et al. [47] introduce an approach to explore causality in medical textual databases. The authors use (medical) domain-specific causal knowledge, such as common causal expressions for depression, schizophrenia, AIDS, and heart diseases, as linguistic clues. Even though these clues play a key role in improving performance, their domain specificity results in a system that can only perform well within the medical domain. Radinsky et al. [79] propose a Pundit algorithm to generate causality pairs from news articles. This

Table 4 Examples of causal patterns from Wikipedia in Ittoo and Bouma [38]

Causal pattern	Linguistic realization
Destroy	"short-circuit in brake wiring *destroyed* the power supply"
Prevent	"message box *prevented* viewer from starting"
Exceed	"breaker voltage *exceeded* allowable limit"
Reduce	"resistor *reduced* voltage output"
Cause	"gray lines *caused* by magnetic influence"
Induce	"bad cable extension might have *induced* the motion problem"
Due to	"replacement of geometry connection cable *due to* wear and tear"
Mar	"cluttered options *mars* console menu"

rule-based approach achieves higher automation than the above two pattern-based systems, thanks to the use of the generalization rule, <Pattern, Constraint, Priority>. The system obtains 70.4% precision on titles taken from news articles spreading over a period of 150 years. However, since the rules only cover obvious causality cases, this system achieves a poor recall metric value of 10.0%. Based on the idea that noun–noun pairs can encode the same semantic relation if they have the same or similar sense collocation, Beamer et al. [5] propose a WordNet-based learning model to capture noun features from WordNet's IS-A backbone. A different approach is followed by Girju et al. [28] that instead of using manually constructed resources, they introduce a model that automatically constructs a set of patterns using a pattern cluster algorithm. Even though [5] outperforms [28] on SemEval-2007 Task 4 by an F-score of 4.8%, it should be noted that the former system has poor portability, since it may be unfeasible to use WordNet and additional resources in other applications or corpora.

5.1.2 Implicit causality

Ittoo and Bouma [38] develop a minimally supervised method to identify three pre-defined types of implicit causality in an iterative way. The first type involves resultative verbal patterns, which include verbs like *increase, reduce, kill, become*. The second type involves patterns that make cause and effect inseparable. The third one involves nonverbal patterns, like *rise in* and *due to*. One innovation of this work is the fact that such defined causal patterns are acquired from Wikipedia, as exemplified in Table 4. An experimental study involving 32,545 documents in the Product Development Customer Service (PD-CS) domain achieved an 85.0% F-score, a result which is comparable to those obtained by state-of-the-art systems. In a study published in 2014, Kang et al. [42] focus on the extraction of adverse drug effects from the ADE corpus. This system consists of a concept identification module that recognizes drugs and adverse effects in sentences, and a knowledge-based module for identifying whether a relationship exists between the recognized concepts. A rule-based NLP module, which consists of a number of rules, is combined with a dictionary-based concept recognition and normalization tool, namely Peregrine, to recognize relevant concepts.

5.1.3 Inter-sentential causality

Khoo et al. [46] propose an approach relying on four kinds of causal links and 2082 causative verbs to construct a set of verbal linguistic CE patterns. A computer program finds all parts of the document that match any of the linguistic patterns, so it is able to identify causal

relations both within a sentence and between adjacent sentences. It achieves an accuracy of 68.0% on a set of 1082 sentences taken from the WSJ newspaper, with many errors caused by complex sentence structure and the lack of inferencing capability from world knowledge. Verbal-linguistic patterns are used to extract relations between mutations in viral genomes (cause) and HIV drugs (effect) in the system of Bui et al. [12]. An initial text retrieval phase sorts out intra- and inter-sentence candidates if there are <mutation, relation, drug> triplets. A subsequent text preprocessing phase simplifies candidate sentences and manually analyzes the existence of causality based on a list of pre-defined keywords. A final relation extraction phase applies eleven causality rules to form linguistic patterns. This system is used in five hospitals to find resistance data in medical literature. However, due to the large number of noun phrases and technical terms, it can be significantly time-consuming and laborious to simplify sentences, which includes removing parenthetical remarks, replacing *known* terms, grouping mutation and drug names, normalizing sentences, and resolving anaphoras.

5.2 Statistical machine learning-based approaches

Statistical machine learning-based approaches require less manual pre-defined patterns than knowledge-based approaches. They usually employ third-party NLP tools (e.g., Spacy [36], Stanford CoreNLP [59], Stanza [76]) to generate a set of features for a given collection of labeled data, and subsequently use ML algorithms (e.g., support vector machine (SVM), maximum entropy (ME), naĭve bayes (NB), and logistic regression (LG)) to perform the relevant classification. In the following paragraphs we explain in more detail how statistical machine learning techniques are used in CE systems.

5.2.1 Explicit Intra-sentential causality

Inspired by a previous work which uses lexico-syntactic patterns to infer causation in a semi-automatic way, in 2003, Girju [26] proposes a model to detect causal relations in a QA system. This model focuses on the most frequent explicit intra-sentential causality patterns, *<NP1, verb, NP2>*, where the verb is a simple causative, and then validates those patterns referring to causation through a set of features based on a decision tree (DT). Blanco et al. [10] identify causality following the *<VerbPhrase, relator, Cause>* pattern only, where relator is one of *because, since, after, as*. The authors use seven kinds of features with a DT on a popular question classification dataset known as TREC [85], and achieve an F-score of 91.3%. Most errors in this system occur when the relator in the pattern is *as* or *after*. This means that the system tends to only perform well when the instances have clear occurrences of the *because* or *since* keywords. Pakray and Gelbukh [69], Sorgente et al. [83], and Zhao et al. [99] evaluate their models on the corpus of SemEval-2010 Task 8. The first two methods identify plausible causality instances based on some common sentence-level features, e.g., contextual, constituent parse, and dependency parse features, and then use DT and Bayesian inference, respectively, to discard non-causal instances. Based on the idea that similar causal connectives have similar ways of expressing causality, Zhao et al. [99] introduce a new causal connective feature, which is collected from the similarity of the sentence's syntactic structure, to divide connectives into difference classes. A Restricted Hidden Naive Bayes (RHNB) is used to process features and the interactions between causal connectives and lexico-syntactic patterns. The performance of these three models, with F-scores of respectively 85.8%, 73.9%, and 85.6%, demonstrates that the identification of appropriate rules and/or features plays a crucial role in machine learning-based frameworks. Kim et al. [48] combine a probabilistic

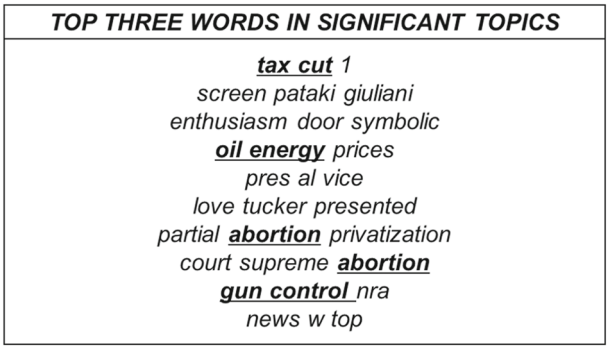

Fig. 2 Sample results in Kim et al. [48]

topic model with time-series causal analysis in order to mine causal topics. It iteratively refines topics, increasing the correlation between discovered topics with time-series data. In one experiment, specific topics that were expected to affect the 2000 presidential election were mined. Figure 2 shows several important issues (e.g., tax cuts, oil energy, and abortion); such topics are typically also cited in the politics-related literature, which shows the efficiency of this model. Lin et al. [57] employ four kinds of features, (production rules, dependency rules, word pairs, and contextual) with an ME classifier. An experiment on PDTB 2.0 indicates that the production rules feature contributes the most to the RE task, followed by word pairs, dependency rules, and contextual features. However, as cause–effect is the most predominant relation in the training set, this model tends to label uncertain relations as causality instances. Thus, it gives this relation high recall but very low precision, leading to an F-score of only 51.0%. Rutherford and Xue [81] employ Brown cluster [11] pairs to represent relations and employ coreference patterns to identify meaningful relations in PDTB 2.0.

5.2.2 Implicit causality

In order to alleviate the shortage of causal connectives, Hidey and McKeown [35] use connectives from PDTB 2.0 AltLex as seed data to identify new alternative lexicalizations from parallel corpora. They train an SVM classifier to process the parallel connectives and lexico-semantic features. The two kinds of features assist the model in achieving the F-score of 75.3%, which is a significant 11.1% improvement over its baseline on AltLex corpus. Inspired by the success of kernel-based machine learning methods on RE, Airola et al. [1] use a dependency-path kernel to extract protein-protein interactions (PPIs) from BioInfer. Each instance is represented by two graphs, one corresponding to the syntactic structure of sentences, and the other to their linear order. In Keskes et al. [44], a ME model is proposed to learn causality in Arabic. Eight linguistic features make significant contributions to identifying implicit relations, like the modality feature to check if a sentence has Arabic modal words based on a manually constructed lexicon. The experiment on newswire stories achieves an F-score of 78.1% and accuracy of 80.6%. However, these rich and complex feature lists rely heavily on NLP tools like the Standard Arabic Morphological Analyzer and Stanford parser. Unfortunately, due to the specific characteristics of different languages, some features that are well extracted may be useless in other languages. Pechsiri et al. [70] utilize verb-pair rules to train NB and SVM to mine implicit causality from Thai texts. WordNet and predefined plant disease information are used to collect the cause and effect verb concepts as a

set of verb-pair rules. The experiment on 3000 agriculture-related sentences obtain precision and recall metrics of 86.0% and 70.0%, respectively. Unlike many other methods that use rich sets of features to represent the input instances, this model focuses on the leverage of task-specific background knowledge, which means that the model can only be applied on a small number of domain-specific texts. In [33] Gurulingappa (the ADE corpus creator) and colleagues train a ME model with simple features, like words in the sentence, to signal the availability of the corpus. Their experimental results, with an F-score of 70.0%, are used as baseline performance values for other systems.

5.2.3 Inter-sentential causality

Marcu and Echihabi [60] utilize lexical pair probability to discriminate causality in inter-sentential forms. They use sentence connecting keywords *Because of* and *Thus* to find candidate sentence pairs and use pre-collected explicit causality nouns, verbs and adverbs to explore causal lexical pairs. Non-causal lexical pairs are obtained from randomly selected sentence pairs. Oh et al. [66] propose a system to explore both intra- and inter-sentential causal relations in a Japanese why-QA system. They utilize regular expressions with explicit keywords in order to identify cue phrases for causality. For each identified cue phrase, the system then extracts three sentences as one causality candidate, including the cue phrase with its left and right sentences. In the process of extracting candidate answers, semantic and syntactic features are used to train a conditional random field (CRF) model to generate cause–effect labels for each word. Finally, to understand chemical induced disease (CID) relations from biomedical articles, Qian and Zhou [77] propose two ME models to extract CID at both the intra- and inter-sentential levels, respectively. They construct training and test instances at inter-sentence level complying with three heuristic rules: (a) only pairs of entities that are not involved in any intra-sentential instance are considered at the inter-sentence level; (b) the sentence distance between two entities should be less than three; (c) if there are multiple entities in the same instance, keep the entity pairs with the shortest distance. The authors then use an ME classifier with lexical features to extract this relationship from a collection of 1500 medical articles.

5.3 Deep learning-based approaches

Neural networks (NNs) are basic algorithms for deep learning (DL). Similarly to a human's neural system, an NN is composed of neurons in three kinds of layers: input, hidden, and output. Each neuron receives input from preceding neurons and produces an output for subsequent neurons. When an NN learns multiple levels of representation from multiple hidden layers, it is said to be a "deep" neural network, and the process is referred to as 'deep learning' [53].

Compared with knowledge-based and statistical ML-based models, deep learning models map words and features into low-dimensional dense vectors, which may alleviate the feature sparsity problem. Furthermore, the use of an attention mechanism to selectively concentrate on relevant aspects, while ignoring others, tends to make deep learning models more effective. The most typical deep learning models include convolutional neural networks (CNNs), recurrent neural networks (RNNs), and variants of the latter like long short-term memory (LSTM) and gated recurrent units (GRU). Later, the introduction of unsupervised pretraining language models (PTMs) like BERT [22], which return contextualized embeddings for each token, significantly improved the performance on many NLP tasks [8]. Both CNNs and RNNs

can be viewed as sequential-based models, which embed semantic and syntactic information in local consecutive word sequences [93]. In comparison, graph-based models, like graph convolutional networks (GCNs) and graph attention networks (GATs), which model a set of points (nodes) and their relationships (edges), have also received the attention of researchers.

In the following paragraphs we discuss how deep learning architectures have been used to solve the CE problem.

5.3.1 Explicit intra-sentential causality

The studies of [51,56,89,91,98] employ different deep learning models in order to extract causality from the SemEval-2010 Task 8 dataset. Xu et al. [91] use LSTM to learn higher-level semantic and syntactic representations along the shortest dependency path (SDP), while Li et al. [56] combine BiLSTM with multi-head self-attention (MHSA) to direct attention to long-range dependencies between words. Cases analysis shows that MHSA significantly improves the performance when the causality distance is greater than 10. Wang et al. [89] propose a multi-level attention-based CNN model to capture entity-specific and relation-specific information. More specifically, an attention pooling layer is used to capture the most useful convolved features of CNN. The studies of [56,89] demonstrate the efficiency of attention, especially of the multi-attention mechanism, in the CE task. Zhang et al. [98] propose a dependency tree-based GCN model to extract relationships. A new tree pruning strategy is applied in order to incorporate relevant information and remove irrelevant context, keeping words that are directly connected to the SDP. Similarly to [91], this technique is based on the definition of the SDP that is used in NLP tools, which will inevitably generate cascading errors. This is the main reason for the F-score of [98] to be lower than [89] by 3.2%. Kyriakakis et al. [51] explore the application of PTMs like BERT and ELMO [72] in the context of CE, by using bidirectional GRU with self-attention (BIGRUATT) as the baseline. The experimental results show that PTMs are only helpful for datasets with hundreds of training examples, and that BIGRUATT reaches a performance plateau when thousands of training instances are available. It should be noted that this finding is inconsistent with the results of other studies, which have shown that PTMs are helpful regardless of the magnitudes of the training sets. The TACRED creators, Zhang et al. [97], combine LSTM with entity position-aware attention to encode both semantic information and global positions of the entities. The ablation studies show that this position-aware mechanism is effective and pushes the F-score up by 3.9%. Ponti and Korhonen [73], Chen et al. [15], Lan et al. [52] focus on causality extraction from the PDTB 2.0 corpus. Ponti and Korhonen [73] develop a feedforward neural network (FNN) model that combines positional and event-related features with basic lexical features to obtain an enriched feature set. Positional features encode the distance between each word to the *cause* and *effect*, while event-related features account for the semantics of the input sentences. Experimental evaluation of performance indicates that positional features have a positive impact on causality identification. Instead of relying on LSTM only, Chen et al. [15] incorporate BiLSTM with GRU in order to capture more complex semantic interactions between text segments. Finally, the method followed in [52] is an attention-based LSTM model that can perform two kinds of representation learning simultaneously: The attention-based module conducts relation representation learning from the interaction between two segments, while a multi-task learning framework uses external corpora to continually improve the performance.

Sentence from the test set	Training	Test	B2	Our proposal
The United States decided to *break* off economic relations with Cuba (which means that they would stop buying things from them).	Causal	Non Causal	Causal	Non Causal
Although Roosevelt had promised *to keep* the United States out of the war, he nevertheless took concrete steps to prepare for war.	Causal	Non Causal	Causal	Non Causal
Mary spent the next 18 years in confinement, but proved too dangerous *to keep* alive, as the Catholic powers in Europe considered her, not Elizabeth, the legitimate ruler of England.	Causal	Non Causal	Causal	Non Causal
Greatly alarmed and with Hitler *making* further demands on the Free City of Danzig, Britain and France guaranteed their support for Polish independence; when Italy conquered Albania in April 1939, the same guarantee was extended to Romania and Greece.	Causal	Non Causal	Causal	Non Causal
They are purely written languages and are often *difficult* to read aloud.	Causal	Non Causal	Causal	Non Causal

Fig. 3 Sample results in Martínez-Cámara et al. [61]

5.3.2 Implicit causality

Martínez-Cámara et al. [61] believe that the use of linguistic features may restrict the ability to represent causality, so they propose an LSTM model incorporating only word embeddings as input. The experimental results obtained on the PDTB 2.0 AltLex corpus show an F-score of 81.9%. Figure 3 shows five examples from this study. For instance, the first example in the figure indicates that the verb *break* mostly has a causal meaning in training examples, but is not a causative verb in the test sentence. It is misclassified by B2 (a baseline model), but correctly classified by the proposed approach. In the study of [14], the authors first incorporate reinforcement learning (RL) to relabel noisy-labeled instances, and then use PCNN, a piece-wise CNN-based model, to iteratively retrain relation extractors with adjusted labels. As the joint entity-relation extraction method can benefit from a close interaction between entities and their relations, Li et al. [55], Wang and Lu [88], Zhao et al. [100] propose joint models for entity and relation extraction from the ADE corpus. Li et al. [55] feed character-level representations, word and part-of-speech (POS) embeddings into a BiLSTM to learn entities and context representations. Another BiLSTM is used to learn the relation representation along with the SDP. Wang and Lu [88] focus on encoding sequence representations and table representations for recognizing entities and their relations, respectively; the two representations then interact with each other attempting to capture task-specific information. The cross-modal attention network (CMAN) in [100] is constructed by stacking two attention units, known as BiLSTM-enhanced self-attention (BSA) unit and BiLSTM-enhanced label-attention (BLA) unit, in order to obtain dense correlations over token and label spaces. Another BLA unit captures token–relation interactions to form the final label-aware token features. The experimental results on ADE show that CMAN achieves state-of-the-art performance with an F-score of 81.1%, surpassing [88] by 1.0%. Two other joint models for ADE extraction can be found in [6,7].

5.3.3 Inter-sentential causality

Taking full advantage of the fact that BiLSTM may alleviate the issue of learning long-range dependencies from a sequence of words, Jin et al. [39] use CNN to capture essential features from input examples, and then utilize BiLSTM to obtain deeper contextual semantic information between cause and effect. Similarly, after extending the annotation of SemEval2010 Task 8 to phrase-level, Dasgupta et al. [21] propose a linguistically informed BiLSTM model

to encode word embeddings with linguistic features. The main reason for the misclassification of causal instances as non-causal instances is that the dependency parser fails to parse the texts correctly, and thus returns improperly syntactic features. Kruengkrai et al. [49] introduce a variant of CNN, called multi-column CNN, to recognize event causalities. Based on the assumption that dependency paths between cause and effect can be viewed as background knowledge, they use a wide range of such paths, regardless of whether cause and effect appear within one sentence or in adjacent sentences, taking web texts as extra input. Within this model, different columns represent different inputs, such as event causality candidates, contextual information, and background knowledge, with each column having its independent convolutional and pooling layers. All the outputs are concatenated into a SoftMax function to perform the classification. The experimental results demonstrate that related background knowledge significantly improves the performance.

6 Systems summary

In the previous section we reviewed 45 systems regarding the different causality forms extracted, i.e., explicit intra-sentential, implicit, and inter-sentential causality, and the different techniques and models employed, i.e., knowledge-based, statistical ML-based, and deep learning-based. Figure 4 contains a statistical summary of the reviewed models. In terms of causality forms, 18.2%, 31.8%, and 50.0% of the 45 systems focus respectively on inter-sentential, implicit and explicit intra-sentential causality. In terms of techniques, 20.4%, 36.4%, and 43.2% of the systems utilize knowledge-based, statistical ML-based, and deep learning-based models. In the next three paragraphs, we separately summarize the advantages and limitations of using each of the three families of techniques.

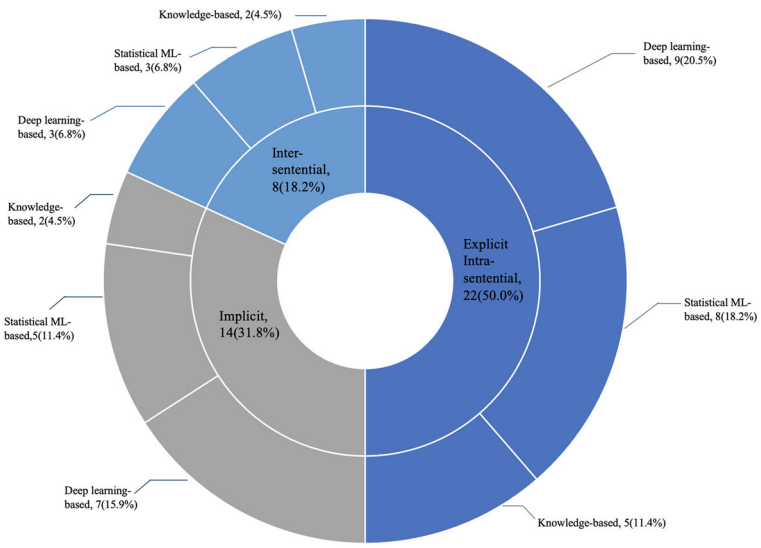

Fig. 4 Statistical summary of the reviewed systems

6.1 Knowledge-based approaches

These approaches rely on the most straightforward methods, using predefined linguistic rules or patterns to detect whether there exist causal relations hidden in the context. Therefore, they can make use of clear keywords in explicit intra-sentential causality as linguistic clues. On the other hand, since explicit connectives are missing in both implicit and inter-sentential causality, these systems require significant effort to prepare complicated clues, especially involving word-level patterns. Causality extraction by pattern or rule matching can perform well in restricted domains; however, since preparing various kinds of clues is a time-consuming task, knowledge-based methods are unsuitable when the consistency of the dataset is poor.

6.2 Statistical machine learning-based approaches

Instead of collecting predefined patterns or rules, traditional ML-based models utilize rich linguistic features or elaborately designed kernels. When the annotated dataset is in the explicit intra-sentential causality form, these systems apply classifiers, with manually or automatically collected features, to remove non-causal instances. When explicit keywords are missing, more complicated features may also need to be prepared. Statistical ML-based approaches usually achieve better performance than knowledge-based systems; however, well-prepared features or kernels may lead to weak portability downstream.

6.3 Deep learning-based approaches

Since deep learning can automatically deduce higher-level information from input vectors and make adjustments to the expected results, these systems are able to focus more on the choice of input features and model architecture, rather than on the preparation of linguistic information. Deep learning-based systems thus have better portability to different applications. However, they require access to larger corpora and more substantial computational resources than the other two techniques.

In conclusion, as illustrated in Fig. 5, specific needs and other contextual aspects should be taken into consideration in order to choose an appropriate method for causality extraction.

Among the reviewed systems, 28 approaches evaluate their models using the benchmark datasets that we introduced in Sect. 3. From Table 5, we can see that deep learning-based methods achieve new state-of-the-art results, and even show substantial improvements in most of the benchmark datasets. For instance, the F-score of Kyriakakis et al. [51] surpasses [83] by 16.7% (on the SemEval-2010 Task 8 corpus), and [100] achieves a higher F-score than [42] by 25.8% (on the ADE corpus). A notable exception is the deep learning-based model of [14], which achieves an F-score of only 49.8% on BioInfer, poorer than [1] by 11.5%. The main reason for this is that the model extracts relations after relabeling noise data iteratively, which alerts us to the fact that model performance is closely related to data quality. On the other hand, performance is not the only criterion for judging or choosing an approach. When models based on the same technology have similar or even the same results, we should also take the local situation and needs into account. For example, the F-score of Li et al. [56] is lower than the F-score of Zhang et al. [98] by 0.2% on SemEval-2010 Task 8. However, it is able to avoid the error propagation that is characteristic of the tree pruning strategy and may have better portability.

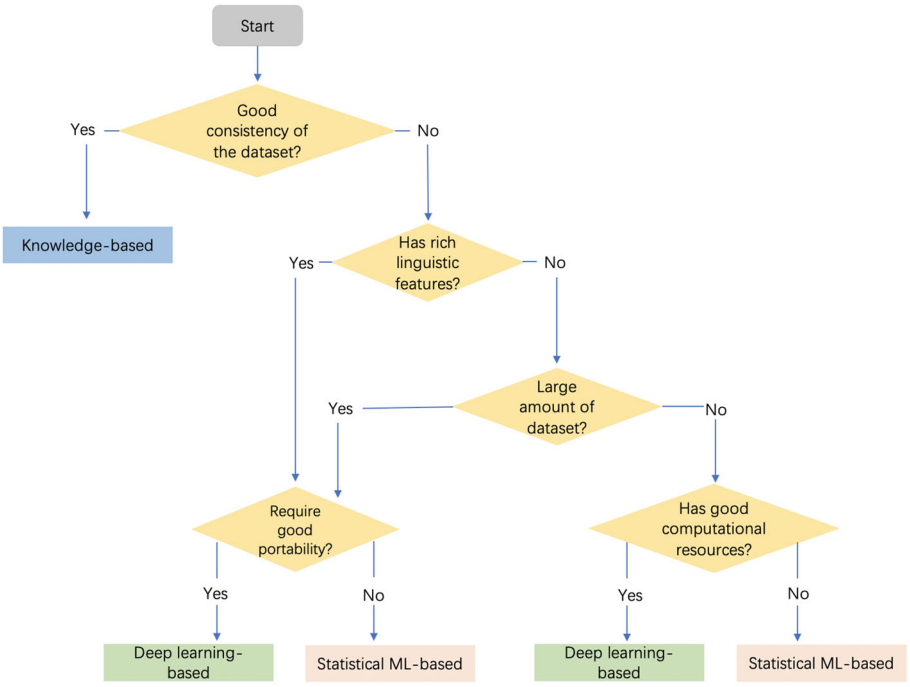

Fig. 5 The process to choose the appropriate technique for causality extraction

The remaining 17 approaches use their own collected datasets or other kinds of publicly available datasets to evaluate their models' performances. We summarize these approaches in Table 6.

7 Open problems and future directions

From the representative systems in Sect. 5, we can know that causal relation extraction has received increasing attention over the past decade. However, it is a non-trivial problem and many challenges remain unsolved, such as the following three problems:

– **Multiple causalities** Most previous CE only focused on one causal pair from an instance, but causality in the real-world literature is more complex. Causal Patterns in Sciences from Harvard Graduate School of Education introduce three common causal patterns[2] as below:

  – (9a) Domino Causality that one cause produces multiple effects.
  – (9b) Relational Causality that two causes work together to produce an effect.
  – (9c) Mutual Causality that cause and effect impact each other simultaneously, or sequentially.

  Like the study of [21], traditional ways to deal the above kinds of multiple causalities is dividing sentence into several sub-sentences that extracted causal pairs separately. This

[2] https://www.cfa.harvard.edu/smg/Website/UCP/causal/causal_types.html.

Table 5 Approaches on benchmark datasets

Dataset	System	Year	Technique	F-score (%)
SemEval-2007	Beamer et al. [5]	2008	Knowledge-based	65.8
Task 4	Girju et al. [28]	2009	Knowledge-based	**70.6**
SemEval-2010	Sorgente et al. [83]	2013	Statistical ML-based	73.9
Task 8	Xu et al. [91]	2015	Deep learning-based	83.7
	Li et al. [56]	2021	Deep learning-based	84.6
	Zhang et al. [98]	2018	Deep learning-based	84.8
	Zhao et al. [99]	2016	Statistical ML-based	85.6
	Pakray and Gelbukh [69]	2014	Statistical ML-based	85.8
	Wang et al. [89]	2016	Deep learning-based	88.0
	Kyriakakis et al. [51]	2019	Deep learning-based	**90.6**
PDTB 2.0	Lin et al. [57]	2009	Statistical ML-based	51.0
	Rutherford and Xue [81]	2014	Statistical ML-based	54.4
	Ponti and Korhonen [73]	2017	Deep learning-based	54.5
	Chen et al. [15]	2016	Deep learning-based	54.8
	Lan et al. [52]	2017	Deep learning-based	**58.9**
PDTB 2.0	Hidey and McKeown [35]	2016	Statistical ML-based	75.3
AltLex	Martínez-Cámara et al. [61]	2017	Deep learning-based	**81.9**
TACRED	Zhang et al. [97]	2017	Deep learning-based	65.4
	Zhang et al. [98]	2018	Deep learning-based	**68.2**
BioInfer	Chen et al. [14]	2020	Deep learning-based	49.8
	Airola et al. [1]	2008	Statistical ML-based	**61.3**
ADE	Kang et al. [42]	2014	Knowledge-based	54.3
	Gurulingappa et al. [33]	2012	Statistical ML-based	70.0
	Li et al. [55]	2017	Deep learning-based	71.4
	Bekoulis et al. [7]	2018	Deep learning-based	74.6
	Bekoulis et al. [6]	2018	Deep learning-based	75.5
	Wang and Lu [88]	2020	Deep learning-based	80.1
	Zhao et al. [100]	2020	Deep learning-based	**81.1**

Bold values indicates the highest F-score in each dataset

method is computationally expensive and cannot take into consideration the dependencies among causality pairs.

The Tag2Triplet algorithm from [56] can extract multiple causal triplets simultaneously. It counts the number and the distribution of each causal tag to judge the tag as simple causality or complexity causality. Afterward, it applies a Cartisian Product of the causal entities to generate possible causal triplets. In addition, [17,87] utilize deep learning with relational reasoning to identify multiple relations in one instance simultaneously.

– **Data deficiency** Typically, for many classification tasks, more than 10 million samples are required to train a deep learning model, so that it can match or exceed human performance [29]. However, just as the size of the four benchmark datasets introduced in Sect. 3 is far from the size of a satisfactory deep learning model, the annotated data in the real world is very specific and small.

Table 6 Approaches on other datasets

Technique	System	Year	Causality form	Dataset	Performance
Knowledge-based	Khoo et al. [46]	1998	Inter-sentential	1082 WSJ sentences	Accuracy (68.0%)
	Khoo et al. [47]	2000	Explicit intra-sentential	130 medical abstracts	Precision (68.0%)
	Garcia et al. [25]	2006	Explicit intra-sentential	Technical texts in French	Precision (85.2%)
	Bui et al. [12]	2010	Inter-sentential	630 medical sentences	F-score (84.0%)
	Radinsky et al. [79]	2012	Explicit intra-sentential	150 years of news articles	Precision (77.8%)
	Ittoo and Bouma [38]	2013	Implicit	32,545 documents in PD-CS	F-score (85.0%)
Statistical ML-based	Marcu and Echihabi [60]	2002	Inter-sentential	BLIPP	Accuracy (87.3%)
	Girju [26]	2003	Explicit intra-sentential	TREC	Precision (73.9%)
	Pechsiri et al. [70]	2006	Implicit	3000 medical sentences in Thai	Pre cision (86.0%)
	Blanco et al. [10]	2008	Explicit intra-sentential	TREC	F-score (91.3%)
	Kim et al. [48]	2013	Explicit intra-sentential	Six months of news articles	t-value (3.87)
	Oh et al. [66]	2013	Inter-sentential	850 Japanese QA examples	F-score (77.0%)
	Keskes et al. [44]	2014	Implicit	90 documents in Arabic	F-score (80.6%)
	Qian and Zhou [77]	2016	Inter-sentential	1500 medical abstracts	Precision (58.3%)
Deep learning-based	Kruengkrai et al. [49]	2017	Inter-sentential	159,350 web sentences	Precision (55.1%)
	Dasgupta et al. [21]	2018	Inter-sentential	Extended semEval-2010 Task 8	F-score (66.0%)
	Jin et al. [39]	2020	Inter-sentential	1986 sentences in Chinese	F-score (82.3%)

Based on the assumption that *any sentence that contains a pair of entities that participate in a known Freebase relation is likely to express that relation*, Mintz et al. [65] introduce the first distant supervision (DS) system for relation extraction, which creates and labels training instances by Freebase as a relation labels resource. However, this method suffers from a large amount of noise labeled data. The survey of [82] introduces methods of addressing the problem of incomplete and wrong labels from DS, like at-least-one models, topic-based models, and pattern correlations models. The very recent research from Huang and Wong [37] proposes a novel way for relation extraction from insufficient labeled data. They first utilize a BiLSTM with attention mechanism to encodes sentences in an unsupervised learning way, the word sequence of entity pairs act as the relation embeddings. Afterward, a random forest classifier is used to learn the relation types from these relation embeddings. This approach of combine unsupervised learning with supervised learning provides us another new idea of solving data deficiency problem in CE task.

– **Document-level causality** Both intra- and inter-sentential causality are at the sentence-level, in real-world scenarios; however, large amounts of causality span multiple sentences, and even in different paragraphs. Unlike being extracted through linguistic cues or features directly, a satisfactory document-level CE requires that the model has strong pattern recognition, logical and common-sense reasoning [18]. All of these aspects need long-term research and exploration.
Zeng et al. [95] introduce a system of combine GCN with relational reasoning to extract relations within a document. They first construct a mention-level GCN to model complex interaction among entities, and then utilize a path reasoning mechanism to infer relations between two entities. This method outperforms the state of the art on the public dataset, DocRED from Yao et al. [94]. Similar approaches can be found in [64,86].

8 Conclusion

Causal relations in natural language text play a key role in clinical decision-making, biomedical knowledge discovery, emergency management, news topic references, etc. Therefore, successful causality extraction from fast-growing unstructured text data is a fundamental task toward constructing a causality knowledge base. In this paper, we conducted a comprehensive review of CE in which we introduced six kinds of benchmark datasets and the evaluation metrics for this task. Afterward, we reviewed existing approaches that use traditional or modern techniques to extract different causality forms. From Sects. 5 and 6, we know that the critical step to extract explicit and implicit causality is to prepare linguistic keywords, patterns, and features, while intra-sentential and inter-sentential causality depend on the way of preparing input instances. Also, we introduced three challenges, which are multiple causalities, data deficiency, and document-level causality extraction, with their potential solutions.

Deep learning provides promising directions for CE tasks. Specifically, domain-related PTMs with graph-based model hold great potential for the two reasons: 1) As the word distributions of general-purpose corpora are quite different with the word distributions of specific domain corpora, the standard PTMs has been shown not to perform well in specialized domains [22]. In contrast, pre-training from scratch on domain-specific corpora, like SciBERT [8], BioBERT [40], and BlueBERT [71], can alleviate this limitation. 2) The study of [32,98] demonstrates the advantage of GCN in complex texts. Thus, we can solve the CE problem by combining domain-specific PTMs with graph models.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10115-022-01665-w.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

References

1. Airola A, Pyysalo S, Björne J, Pahikkala T, Ginter F, Salakoski T (2008) All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. BMC Bioinform 9(11):S2. https://doi.org/10.1186/1471-2105-9-S11-S2
2. Asghar N (2016) Automatic extraction of causal relations from natural language texts: a comprehensive survey. arXiv preprint arXiv:1605.07895
3. Balashankar A, Chakraborty S, Fraiberger S, Subramanian L (2019) Identifying predictive causal factors from news streams. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), association for computational linguistics, Hong Kong, China, pp 2338–234. https://doi.org/10.18653/v1/D19-1238
4. Barik B, Marsi E, Ozturk P (2016) Event causality extraction from natural science literature. Res Comput Sci 117:97–107. https://doi.org/10.13053/rcs-117-1-8
5. Beamer B, Rozovskaya A, Girju R (2008) Automatic semantic relation extraction with multiple boundary generation. In: Proceedings of the 23rd national conference on artificial intelligence. AAAI Press, Chicago, Illinois, pp 824–829
6. Bekoulis G, Deleu J, Demeester T, Develder C (2018a) Adversarial training for multi-context joint entity and relation extraction. In: Proceedings of the 2018 conference on empirical methods in natural language processing, association for computational linguistics. Brussels, Belgium, pp 2830–2836. https://doi.org/10.18653/v1/D18-1307
7. Bekoulis G, Deleu J, Demeester T, Develder C (2018) Joint entity recognition and relation extraction as a multi-head selection problem. Expert Syst Appl 114:34–45. https://doi.org/10.1016/j.eswa.2018.07.032
8. Beltagy I, Lo K, Cohan A (2019) Scibert: pretrained language model for scientific text. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, Association for Computational Linguistics, Hong Kong, China, pp 3615–3620. https://doi.org/10.18653/v1/D19-1371
9. Bethard S, Martin JH (2008) Learning semantic links from a corpus of parallel temporal and causal relations. In: Proceedings of ACL-08: HLT, short papers, association for computational linguistics. Columbus, Ohio, pp 177–180
10. Blanco E, Castell N, Moldovan D (2008) Causal relation extraction. In: Proceedings of the international conference on language resources and evaluation. Marrakech, Morocco, pp 310–313
11. Brown PF, Della Pietra VJ, deSouza PV, Lai JC, Mercer RL (1992) Class-based *n*-gram models of natural language. Comput Ling 18(4):467–480. https://aclanthology.org/J92-4003
12. Bui QC, Nuallain OB, Boucher CA, Sloot PM (2010) Extracting causal relations on hiv drug resistance from literature. BMC Bioinform 11(1):101–110. https://doi.org/10.1186/1471-2105-11-101
13. Chang DS, Choi KS (2006) Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities. Inf Process Manage 42(3):662–678. https://doi.org/10.1016/j.ipm.2005.04.004
14. Chen D, Li Y, Lei K, Shen Y (2020) Relabel the noise: joint extraction of entities and relations via cooperative multiagents. In: Jurafsky D, Chai J, Schluter N, Tetreault JR (eds) Proceedings of the 58th annual meeting of the association for computational linguistics, ACL 2020. Online, July 5-10, 2020, Association for Computational Linguistics, pp 5940–5950. https://doi.org/10.18653/v1/2020.acl-main.527

15. Chen J, Zhang Q, Liu P, Qiu X, Huang X (2016) Implicit discourse relation detection via a deep architecture with gated relevance network. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers),.Association for Computational Linguistics, Berlin, Germany, pp 1726–1735. https://doi.org/10.18653/v1/P16-1163
16. Chicco D, Jurman G (2020) The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC Genomics 21(1):6. https://doi.org/10.1186/s12864-019-6413-7
17. Christopoulou F, Miwa M, Ananiadou S (2018) A walk-based model on entity graphs for relation extraction. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: short papers). Association for Computational Linguistics, Melbourne, Australia, pp 81–88. https://doi.org/10.18653/v1/P18-2014
18. Christopoulou F, Miwa M, Ananiadou S (2019) Connecting the dots: document-level neural relation extraction with edge-oriented graphs. 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). Association for Computational Linguistics, pp 4927–4938. https://doi.org/10.18653/v1/D19-1498
19. Cole SV, Royal MD, Valtorta MG, Huhns MN, Bowles JB (2006) A lightweight tool for automatically extracting causal relationships from text. Proc IEEE SoutheastCon 2006:125–129. https://doi.org/10.1109/second.2006.1629336
20. Cowie J, Lehnert W (1996) Information extraction. Commun ACM 39(1):80–91. https://doi.org/10.1145/234173.234209
21. Dasgupta T, Saha R, Dey L, 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. Association for Computational Linguistics, Melbourne, Australia, pp 306–316. https://doi.org/10.18653/v1/W18-5035
22. Devlin J, Chang MW, Lee K, Toutanova K (2019) Bert: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT, pp 4171–4186. https://doi.org/10.18653/v1/N19-1423
23. Dong GF, Zheng L, Huang SH, Gao J, Zuo YC (2021) Amino acid reduction can help to improve the identification of antimicrobial peptides and their functional activities. Front Genet 12:549. https://doi.org/10.3389/fgene.2021.669328
24. Ertekin S, Huang J, Bottou L, Giles L (2007) Learning on the border: active learning in imbalanced data classification. In: Proceedings of the sixteenth ACM conference on conference on information and knowledge management. Association for Computing Machinery, New York, NY, USA, CIKM '07, pp 127–136. https://doi.org/10.1145/1321440.1321461
25. Garcia D, EDF-DER, IMA-TIEM (2006) COATIS, an NLP system to locate expressions of actions connected by causality links, vol 1319. Springer, pp 347–352 (chap BFb0026799). https://doi.org/10.1007/BFb0026799
26. Girju R (2003) Automatic detection of causal relations for question answering. In: Proceedings of the ACL 2003 workshop on multilingual summarization and question answering, Association for Computational Linguistics, USA, MultiSumQA '03, vol 12, pp 76–83. https://doi.org/10.3115/1119312.1119322
27. Girju R, Nakov P, Nastase V, Szpakowicz S, Turney P, Yuret D (2007) Semeval-2007 task 04: classification of semantic relations between nominals. In: Proceedings of the 4th international workshop on semantic evaluations. Association for Computational Linguistics, USA, SemEval '07, pp 13–18
28. Girju R, Nakov P, Nastase V, Szpakowicz S, Turney P, Yuret D (2009) Classification of semantic relations between nominals. Lang Resour Eval 43(2):105–121. https://doi.org/10.1007/s10579-009-9083-2
29. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press
30. Granger CWJ (1969) Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37(3):424–438. https://doi.org/10.2307/1912791
31. Gu Q, Zhu L, Cai Z (2009) Evaluation measures of the classification performance of imbalanced data sets. In: Computational intelligence and intelligent systems, communications in computer and information science, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 461–471. https://doi.org/10.1007/978-3-642-04962-0_53
32. Guo Z, Zhang Y, Lu W (2019) Attention guided graph convolutional networks for relation extraction. In: Proceedings of the 57th annual meeting of the association for computational linguistics, Association for Computational Linguistics, Florence, Italy, pp 241–251. https://doi.org/10.18653/v1/P19-1024
33. Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L (2012) Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. J Biomed Inform 45(5):885–892. https://doi.org/10.1016/j.jbi.2012.04.008
34. Hendrickx I, Kim SN, Kozareva Z, Nakov P, Ó Séaghdha D, Padó S, Pennacchiotti M, Romano L, Szpakowicz S (2010) SemEval-2010 task 8: Multi-way classification of semantic relations between pairs

of nominals. In: Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Uppsala, Sweden, pp 33–38
35. Hidey C, McKeown K (2016) Identifying causal relations using parallel Wikipedia articles. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers). Association for computational linguistics, Berlin, Germany, pp 1424–1433. https://doi.org/10.18653/v1/P16-1135
36. Honnibal M, Montani I (2017) spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing (to appear)
37. Huang H, Wong R (2020) Deep embedding for relation extraction on insufficient labelled data. In: 2020 international joint conference on neural networks (IJCNN), pp 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207554
38. Ittoo A, Bouma G (2013) Minimally-supervised learning of domain-specific causal relations using an open-domain corpus as knowledge base. Data Knowl Eng 88:142–163. https://doi.org/10.1016/j.datak.2013.08.004
39. Jin X, Wang X, Luo X, Huang S, Gu S (2020) Inter-sentence and implicit causality extraction from chinese corpus. Pacific-Asia Conf Knowl Discov Data Min Springer 12084:739–751. https://doi.org/10.1007/978-3-030-47426-3_57
40. Jinhyuk L, Wonjin Y, Kim, (2019) Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234–1240. https://doi.org/10.1093/bioinformatics/btz682
41. Kadir RA, Bokharaeian B (2013) Overview of biomedical relations extraction using hybrid rule-based approaches. J Indus Intell Inf 1(3):169–173. https://doi.org/10.12720/jiii.1.3.169-173
42. Kang N, Singh B, Bui QC, Afzal Z, van Mulligen EM, Kors J (2014) Knowledge-based extraction of adverse drug events from biomedical text. BMC Bioinform 15:64. https://doi.org/10.1186/1471-2105-15-64
43. Karagiannopoulos MG, Anyfantis DS, Kotsiantis SB, Pintelas PE (2007) Local cost sensitive learning for handling imbalanced data sets. In: 2007 Mediterranean conference on control automation, pp 1–6. https://doi.org/10.1109/MED.2007.4433808
44. Keskes I, Zitoune FB, Belguith L (2014) Learning explicit and implicit arabic discourse relations. J King Saud Univ Comput Inf Sci Arch 26:398–416. https://doi.org/10.1016/j.jksuci.2014.06.001
45. Khoo C, Chan S, Niu Y (2002) The many facets of the cause-effect relation. The Semantics of Relationships, pp 51–70. https://doi.org/10.1007/978-94-017-0073-3_4
46. Khoo CSG, Kornfilt J, Oddy RN, Myaeng SH (1998) Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. Literary Ling Comput 13(4):177–186. https://doi.org/10.1093/llc/13.4.177
47. Khoo CSG, Chan S, Niu Y (2000) Extracting causal knowledge from a medical database using graphical patterns. In: Proceedings of the 38th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Hong Kong, pp 336–343. https://doi.org/10.3115/1075218.1075261
48. Kim H, Castellanos M, Hsu M, Zhai C, Rietz T, Diermeier D (2013) Mining causal topics in text data: iterative topic modeling with time series feedback. In: CIKM 2013—proceedings of the 22nd ACM international conference on information and knowledge management. International Conference on Information and Knowledge Management, Proceedings, pp 885–890. https://doi.org/10.1145/2505515.2505612
49. Kruengkrai C, Torisawa K, Hashimoto C, Kloetzer J, Oh JH, Tanaka M (2017) Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks. In: Proceedings of the thirty-first AAAI conference on artificial intelligence. AAAI Press, pp 3466-3473
50. Kubat M, Matwin S (1997) Addressing the curse of imbalanced training sets: one-sided selection. In: Fisher DH (ed) Proceedings of the fourteenth international conference on machine learning (ICML 1997). Nashville, Tennessee, USA, July 8–12, 1997. Morgan Kaufmann, pp 179–186
51. Kyriakakis M, Androutsopoulos I, Saudabayev A, Ginés i Ametllé J (2019) Transfer learning for causal sentence detection. In: Proceedings of the 18th BioNLP workshop and shared task. Association for Computational Linguistics, Florence, Italy, pp 292–297. https://doi.org/10.18653/v1/W19-5031
52. Lan M, Wang J, Wu Y, Niu ZY, Wang H (2017) Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In: Proceedings of the 2017 conference on empirical methods in natural language processing. Association for Computational Linguistics, Copenhagen, Denmark, pp 1299–1308. https://doi.org/10.18653/v1/D17-1134
53. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444. https://doi.org/10.1038/nature14539

54. Li A, Deng Y, Tan Y, Chen M (2021) A transfer learning-based approach for lysine propionylation prediction. Front Physiol 12:452. https://doi.org/10.3389/fphys.2021.658633
55. Li F, Zhang M, Fu G, Ji D (2017) A neural joint model for entity and relation extraction from biomedical text. BMC Bioinform 18. https://doi.org/10.1186/s12859-017-1609-9
56. Li Z, Li Q, Zou X, Ren J (2021) Causality extraction based on self-attentive bilstm-crf with transferred embeddings. Neurocomputing 423:207–219. https://doi.org/10.1016/j.neucom.2020.08.078
57. Lin Z, Kan MY, Ng HT (2009) Recognizing implicit discourse relations in the Penn Discourse Treebank. In: Proceedings of the 2009 conference on empirical methods in natural language processing. Association for Computational Linguistics, Singapore, pp 343–351
58. Manavalan B, Shin TH, Kim MO, Lee G (2018) Pip-el: a new ensemble learning method for improved proinflammatory peptide predictions. Front Immunol 9:1783. https://doi.org/10.3389/fimmu.2018.01783
59. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D (2014) The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, association for computational linguistics. Baltimore, Maryland, pp 55–60. https://doi.org/10.3115/v1/P14-5010
60. Marcu D, Echihabi A (2002) An unsupervised approach to recognizing discourse relations. In: Proceedings of the 40th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp 368–375. https://doi.org/10.3115/1073083.1073145
61. Martínez-Cámara E, Shwartz V, Gurevych I, Dagan I (2017) Neural disambiguation of causal lexical markers based on context. In: IWCS 2017—12th international conference on computational semantics—short papers
62. Matthews B (1975) Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochimica et Biophysica Acta (BBA) Prot Struct 405(2):442–451. https://doi.org/10.1016/0005-2795(75)90109-9
63. Mihaila C, Ananiadou S (2014) Semi-supervised learning of causal relations in biomedical scientific discourse. Biomed Eng Online 13(2):1–24. https://doi.org/10.1186/1475-925X-13-S2-S1
64. Minh Tran H, Nguyen MT, Nguyen TH (2020) The dots have their values: exploiting the node-edge connections in graph-based neural models for document-level relation extraction. In: Findings of the association for computational linguistics: EMNLP 2020. Association for Computational Linguistics, Online, pp 4561–4567. https://doi.org/10.18653/v1/2020.findings-emnlp.409
65. Mintz M, Bills S, Snow R, Jurafsky D (2009) Distant supervision for relation extraction without labeled data. In: Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP. Association for Computational Linguistics, Suntec, Singapore, pp 1003–1011
66. Oh JH, Torisawa K, Hashimoto C, Sano M, De Saeger S, Ohtake K (2013) Why-question answering using intra-and inter-sentential causal relations. In: ACL 2013—51st annual meeting of the association for computational linguistics. Proceedings of the Conference, Sofia, Bulgaria, vol 1, pp 1733–1743
67. Oh JH, Torisawa K, Hashimoto C, Iida R, Tanaka M, Kloetzer J (2016) A semi-supervised learning approach to why-question answering. In: Proceedings of the thirtieth AAAI conference on artificial intelligence. AAAI Press, AAAI'16, pp 3022–3029
68. Oh JH, Torisawa K, Kruengkrai C, Iida R, Kloetzer J (2017) Multi-column convolutional neural networks with causality-attention for why-question answering. In: Proceedings of the Tenth ACM international conference on web search and data mining, pp 415–424. https://doi.org/10.1145/3018661.3018737
69. Pakray P, Gelbukh A (2014) An open domain causal relation detection from paired nominal. In: 13th Mexican international conference on artificial intelligence (MICAI-2014). Nature-Inspired Computation and Machine Learning, Chiapas, Mexico, vol 8857, pp 261–271. https://doi.org/10.1007/978-3-319-13650-9_24
70. Pechsiri C, Kawtrakul A, Piriyakul R (2006) Mining causality knowledge from textual data. In: Proceedings of the 24th IASTED international conference on artificial intelligence and applications. ACTA Press, USA, AIA'06, pp 85–90
71. Peng, Yifan, Yan, Shankai (2019) Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. In: Proceedings of the BioNLP 2019 workshop. Association for Computational Linguistics, Florence, Italy, pp 58–65
72. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, 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, vol 1 (long papers). Association for Computational Linguistics, New Orleans, Louisiana, pp 2227–2237. https://doi.org/10.18653/v1/N18-1202

73. Ponti EM, 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. Association for Computational Linguistics, Valencia, Spain, pp 25–30. https://doi.org/10.18653/v1/W17-0903
74. Prasad R, Miltsakaki E, Dinesh N, Lee A, Joshi A (2007) The penn discourse treebank 2.0 annotation manual. IRCS technical reports series 203 Philadelphia: University of Pennsylvania ScholarlyCommons, p 105
75. Pyysalo S, Ginter F, Heimonen J, Björne J, Boberg J, Järvinen J, Salakoski T (2007) Bioinfer: a corpus for information extraction in the biomedical domain. BMC Bioinform 8:50. https://doi.org/10.1186/1471-2105-8-50
76. Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD (2020) Stanza: a python natural language processing toolkit for many human languages. In: Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations. Association for Computational Linguistics, Online, pp 101–108. https://doi.org/10.18653/v1/2020.acl-demos.14
77. Qian L, Zhou G (2016) Chemical-induced disease relation extraction with various linguistic features. Database 2016:baw042. https://doi.org/10.1093/database/baw042
78. Qiu J, Xu L, Zhai J, Luo L (2017) Extracting causal relations from emergency cases based on conditional random fields. Procedia Comput Sci 112(C):1623–1632. https://doi.org/10.1016/j.procs.2017.08.252
79. Radinsky K, Davidovich S, Markovitch S (2012) Learning causality for news events prediction. WWW'12—proceedings of the 21st annual conference on world wide web, pp 909–918. https://doi.org/10.1145/2187836.2187958
80. Rink B, Bejan C, Harabagiu S (2010) Learning textual graph patterns to detect causal event relations. In: Proceedings of the 23rd international Florida artificial intelligence research society conference, pp 265–270
81. Rutherford A, Xue N (2014) Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In: Proceedings of the 14th conference of the European chapter of the association for computational linguistics. Association for Computational Linguistics, Gothenburg, Sweden, pp 645–654. https://doi.org/10.3115/v1/E14-1068
82. Smirnova A, Cudré-Mauroux P (2018) Relation extraction using distant supervision: a survey. ACM Comput Surv 51(5). https://doi.org/10.1145/3241741
83. Sorgente A, Vettigli G, Mele F (2013) Automatic extraction of cause-effect relations in natural language text. DART@ AI* IA 1109:37–48
84. Su CT, Hsiao YH (2007) An evaluation of the robustness of mts for imbalanced data. IEEE Trans Knowl Data Eng 19(10):1321–1332. https://doi.org/10.1109/TKDE.2007.190623
85. Voorhees E (2001) Overview of the trec-9 question answering track. Overview of the TREC-9 Question Answering Track, pp 71–80
86. Wang D, Hu W, Cao E, Sun W (2020) Global-to-local neural networks for document-level relation extraction. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Online, pp 3711–3721. https://doi.org/10.18653/v1/2020.emnlp-main.303
87. Wang H, Tan M, Yu M, Chang S, Wang D, Xu K, Guo X, Potdar S (2019) Extracting multiple-relations in one-pass with pre-trained transformers. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Florence, Italy, pp 1371–1377. https://doi.org/10.18653/v1/P19-1132
88. Wang J, Lu W (2020) Two are better than one: Joint entity and relation extraction with table-sequence encoders. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Online, pp 1706–1721. https://doi.org/10.18653/v1/2020.emnlp-main.133
89. Wang L, Cao Z, de Melo G, Liu Z (2016) Relation classification via multi-level attention CNNs. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers). Association for Computational Linguistics, Berlin, Germany, pp 1298–1307. https://doi.org/10.18653/v1/P16-1123
90. Wu R, Yao Y, Han X, Xie R, Liu Z, Lin F, Lin L, Sun M (2019) Open relation extraction: relational knowledge transfer from supervised data to unsupervised data. 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). Association for Computational Linguistics, Hong Kong, China, pp 219–228. https://doi.org/10.18653/v1/D19-1021
91. Xu Y, Mou L, Li G, Chen Y, Peng H, Jin Z (2015) Classifying relations via long short term memory networks along shortest dependency paths. In: Proceedings of the 2015 conference on empirical methods

in natural language processing. Association for Computational Linguistics, Lisbon, Portugal, pp 1785–1794. https://doi.org/10.18653/v1/D15-1206
92. Yang X, Mao K (2014) Multi level causal relation identification using extended features. Expert Syst Appl 41(16):7171–7181. https://doi.org/10.1016/j.eswa.2014.05.044
93. Yao L, Mao C, Luo Y (2019a) Graph convolutional networks for text classification. In: In 33rd AAAI conference on artificial intelligence, pp 7370–7377
94. Yao Y, Ye D, Li P, Han X, Lin Y, Liu Z, Liu Z, Huang L, Zhou J, Sun M (2019b) DocRED: a large-scale document-level relation extraction dataset. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Florence, Italy, pp 764–777. https://doi.org/10.18653/v1/P19-1074
95. Zeng S, Xu R, Chang B, Li L (2020) Double graph based reasoning for document-level relation extraction. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Online, pp 1630–1640. https://doi.org/10.18653/v1/2020.emnlp-main.127
96. Zhang Q, Chen M, Liu L (2017) A review on entity relation extraction. In: 2017 second international conference on mechanical, control and computer engineering (ICMCCE), vol 1, pp 178–183. https://doi.org/10.1109/ICMCCE.2017.14
97. Zhang Y, Zhong V, Chen D, Angeli G, Manning CD (2017) Position-aware attention and supervised data improve slot filling. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, pp 35–45. https://doi.org/10.18653/v1/D17-1004
98. Zhang Y, Qi P, Manning CD (2018) Graph convolution over pruned dependency trees improves relation extraction. In: Proceedings of the 2018 conference on empirical methods in natural language processing. Association for Computational Linguistics, Brussels, Belgium, pp 2205–2215. https://doi.org/10.18653/v1/D18-1244
99. Zhao S, Liu T, Zhao S, Chen Y, Nie JY (2016) Event causality extraction based on connectives analysis. Neurocomputing 173:1943–1950. https://doi.org/10.1016/j.neucom.2015.09.066
100. Zhao S, Hu M, Cai Z, Liu F (2020) Modeling dense cross-modal interactions for joint entity-relation extraction. 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 4032–4038. https://doi.org/10.24963/ijcai.2020/558
101. Zhou D, Zhong D (2014) Biomedical relation extraction: from binary to complex. Comput Math Methods Med 24:298–473. https://doi.org/10.1155/2014/298473

Jie Yang is a Ph.D. student in Computer Science at the University of Sydney in Australia, advised by Dr. Soyeon Caren Han and Dr. Josiah Poon. Her research interests include causality extraction and biomedical relation extraction.

Soyeon Caren Han is a lecturer at School of Computer Science, University of Sydney, teaching and researching on Natural Language Processing and Artificial Intelligence. During her Ph.D., until 2016, Dr. Han had introduced a novel artificial intelligence-based architecture that enables integrating human expertise and machine learning models. These discoveries led her to successfully secure more than 5 million dollars international/national grants. Caren has been working on proposing deep learning-based natural language processing and text mining algorithms to solve multi-modality. Her models were recognised in various international top-tier conferences, including International Conference on Neural Information Processing (Best Paper Award), and International Conference on Computational Linguistic (Best Area Paper Award).

Josiah Poon is a senior lecturer at School of Computer Science, University of Sydney. Dr. Poon uses traditional machine learning techniques paying particular attention to learning from imbalanced datasets, short string text classification, and the data complexity analysis. Recently, he created a Natural Language Processing (NLP) group at the University of Sydney that uses deep learning techniques by investigating various problems, including chatbot technology, generating pathology report from X-rays and explainable AI in documents. He has coordinated a multidisciplinary team consisting of computer scientists, pharmacists, western medicine & traditional Chinese medicine researchers and practitioners since 2007. He co-leads a joint big-data laboratory for integrative medicine (Acclaim) established between the University of Sydney and the Chinese University of Hong Kong to study medical/health problem using computational tools.