A Survey on Extraction of Causal Relations from Natural Language Text

Jie Yang[a], Soyeon Caren Han[a,*], Josiah Poon[a]

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

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.

*This is to indicate the corresponding author.
Email addresses: jyan4704@uni.sydney.edu.au (Jie Yang), caren.han@sydney.edu.au (Soyeon Caren Han), josiah.poon@sydney.edu.au (Josiah Poon)

Lastly, we highlight existing open challenges with their potential directions.

Keywords: causality extraction, explict intra-sentential causality, implicit causality, inter-sentential causality, deep learning

---

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 [1]. 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 medical literature can support constructing a knowledge graph, which can assisting 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 [2, 3, 4, 5, 6], biomedical [7, 8, 9, 10], emergency management [11], etc.

The task of CE focuses on developing systems for identifying cause-effect relations between pairs of labeled nouns from text [12]. From the aspect of techniques, as shown in Figure 1, there has been a considerable body of CE systems that can be divided into three groups: knowledge-based approaches, statistical machine learning-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) causaltive adverbs and adjectives [13]. Implicit causality means explicit causal valence are replaced by ambiguous connectives, as the connectives in Table 1, or even without any connectives. 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 [14, 7, 15, 16, 9, 17] identify causality in the basic levels, which are explicit and/or intra-sentential forms. 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.

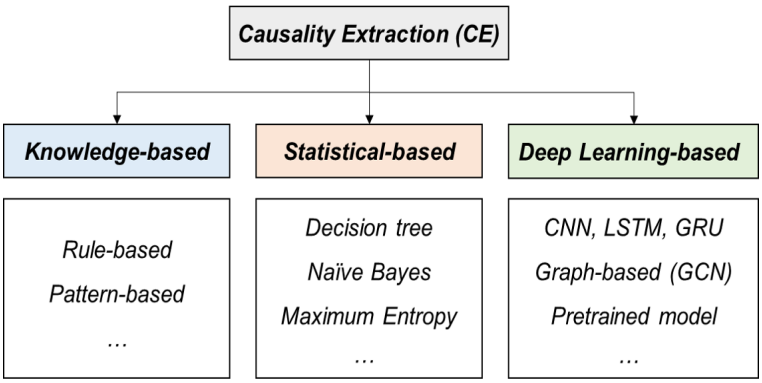

Fig. 1. Taxonomy of techniques

The rest of the article is structured as follows. We review in detail of previous surveys in Section 2. The benchmark datasets and evaluation metrics for CE system are presented in Section 3 and Section 4, respectively. Then, we survey

Table 1

Examples for implicity causality.

Relators	**Sentences**	**Labels**
after	Bischoff in a round table discussion claimed he fired Austin *after* he refused to do a taping in Atlanta. [18]	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. [18]	Non-causal
as	There was no debate *as* the Senate passed the bill on to the House. [19]	Causal
as	It has a fixed time, *as* collectors well know. [19]	Non-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*>

representative CE systems in Section 5. We propose three open problems of the CE task with their potential solutions in Section 6, and the conclusion of this paper are in Section 7.

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 scales up to RE literature reviews. Examples include the generalized survey [20], detailed analyses of RE in the biomedical domain [21, 22], and a survey about the application of distant supervision on RE [23]. 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 [24]. Also, RE faces the challenge of extracting relations on open-domain corpora, that is, the relation types may not be pre-defined [25], while the target of CE is clear and there are no new relation types.

More than a decade ago, Asghar [26] separates CE applications into non-statistical techniques, and statistical & machine learning techniques. Besides reviewing previous approaches, another contribution of Asghar's survey 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 provides 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., [27, 28, 29, 30])

have the capability to explore implicit relations. In the same year, Barik et al. [31] categorize existing CE approaches into four groups: CE using handcrafted patterns, CE using semi-automatic causal patterns, CE using supervised learning, and statistical methods. From their point of view, instead of using manually linguistic clues and domain knowledge, semi-automatic learning acquire lexicon-syntactic patterns from a larger corpus automatically. Then, these patterns are used to identify in-domain causal relations or evaluate causal patterns in a semi-automated 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 Section 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 and related works in Table 3.

- **SemEval-2007 Task 4:** It is part of SemEval (Semantic Evaluation), the 4^{th} edition of the semantic evaluation event [32]. 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% positive data, and 80 test with 51% 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 pre-processing. 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, that 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 [33]. Within the 10,717 annotated examples, there are 1,003 training with 13% positive data, and 328 test with 12% positive data. This small sample amount and imbalanced condition are the major limitations of this dataset.

- **PDTB 2.0:** The second release of the penn discourse treebank (PDTB) dataset from Prasad et al. [34] is the largest annotated corpus of discourse relations. It includes 72,135 non-causal and 9,190 causal examples from 2,312 Wall Street Journal (WSJ) articles. Specifically, there is a type of implicit connectives in the dataset known as AltLex (Alternative lexicalization), which is an open class of markers and potential infinite [35]. However, the authors store PDTB in a complex way, and researchers need to use tools to convert it into easy-to-operate files. Meanwhile, because of the unmarked entities, approaches that need to utilize entities information are unavailable on this corpus.

- **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 2009-2014 [36]. 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.

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

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 filed of biomedical.

- **BioInfer:** Pyysalo et al. [37] introduce an annotated corpus, BioInfer (Bio Information Extraction Resource), which contains 1,100 sentences with the relations of genes, proteins, and RNA from biomedical publications. There are 2,662 relations in the 1,100 sentences, of these 1,461 (55%) are causal-effect. The original data is collected in detail in the XML form, includes sentence with entity markup. Therefore, use this corpus is a matter of simple XML-to-JSON or XML-to-XML transformation.

- **ADE:** The corresponding ADE task aims to extract two entities (drugs and diseases) and relations about drugs with their adverse effects (ADEs) [38, 39]. Dataset in the task is collected from 1,644 PubMed abstracts, in which 6,821 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 [40, 41], only use the 6,821 sentences in the experiments.

4. Evaluation metrics

good background knowledge
clear explanation

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

- Precision = TP/(TP + FP)

[1] https://sites.google.com/site/semeval2007task4/data
[2] https://github.com/sahitya0000/Relation-Classification/tree/master/corpus
[3] https://catalog.ldc.upenn.edu/LDC2008T05
[4] https://catalog.ldc.upenn.edu/LDC2018T24
[5] http://mars.cs.utu.fi/BioInfer/?q=download
[6] https://sites.google.com/site/adecorpus/

Table 3

Benchmark datasets.

Datasets	Published Years	Causality Sizes	Sources	Availability	Related Works
SemEval-2007 Task 4	2007	220	Wikipedia	Publicly available [1]	[12, 42]
SemEval-2010 Task 8	2010	1,331	Wikipedia	Publicly available[2]	[43, 44, 28, 45, 17, 46, 47, 48, 49, 50]
PDTB 2.0	2018	9,190	WSJ	License required[3]	[51, 52, 53, 18, 4, 54, 35, 55]
TACRED	2018	269	Newswire, Web	License required[4]	[36, 44, 47]
BioInfer	2007	1,461	PubMed	Publicly available[5]	[56, 57]
ADE	2012	6,821	PubMed	Publicly avaiable[6]	[38, 58, 39, 59, 60]

- Recall = TP/(TP + FN)
- F-score = 2 * Precision * Recall/(Precision + Recall)
- Accuracy = (TP + TN)/(TP + FP + TN + FN)

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, FN (false negative) is the number of non-causal pairs that are identified as causal pairs.

The entity labelling metrics also applied to evaluate the models. For example, Khoo et al. [2] use average precision and recall to judge whether the model can identifying both the boundary of cause and effect. Dasgupta et al. [43] compare F-score of labeling "C"(cause), "E"(effect), "CC"(causal connectives) and "N"(None) tags with baseline models. Compared with the relation extrac-

tion 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 [61, 62, 63] 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. [64] 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 [65] 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 sort out existing methods of causality extraction based on their techniques as well as the causality forms.

5.1. Knowledge-based approaches

Knowledge-based CE systems can be divided into pattern-based approaches and rule-based approaches. Some of the pattern-based systems express linguistic patterns by 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 analysis, like lexico-semantic and syntactic analysis. These structure analysis lead to performance improvements as well as the chance to extract implicit causality. Similar to pattern-based approaches, some rule-based systems rely on a set of patterns or templates to identify candidate causality directly. In contrast, some rule-based systems utilize a set of procedures or heuristic

algorithms on the syntactic structure of sentences to explore causality. In the following paragraphs, we will introduce how previous systems utilize knowledge-based techniques to extract causality in different forms.

Explicit Intra-sentential Causality Garcia et al. [14] and Khoo et al. [7] use patterns to identify explicitly expressed causal relations within a single sentence. The tool from Garcia et al., COATIS, extracts causality from French texts through lexico-syntactic patterns with 23 explicit causal verbs, like *provoke, disturb, result, lead to*. Due to special attention to the syntactic positions of causal verbs and their surrounding noun phrases, COATIS achieves a reasonably precision of 85%. Even though it only can works on a small amount of French text, it shows how to implement a domain-independent CE system via pre-defined patterns. Khoo et al. [7] introduce an approach to explore causality from medical textual databases. They use medical-specific causal knowledge, like common causal expressions for depression, schizophrenia, AIDs and heart disease, as linguistic clues. Even though these clues play a key role in improved performance, their domain specificity make the system only can work well in medical domain. Radinsky et al. [15] propose a Pundit algorithm to generalize causality pairs from news articles. This rule-based approach is more automatic than the above two pattern-based systems because of the generalization rule, $< Pattern, Constraint, Priority >$. The system is evaluated on titles of 150 years news articles, and achieves 10% recall and 70% precision. However, due to the rules only can cover obvious causality cases, this system achieves very pool recall performance.

Implicit Causality In the study of [66], Ittoo and Bouma 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 non-verbal patterns, like *rise in, due to*. One innovation of the study is acquiring such defined causal patterns from Wikipedia, as the examples in Table 4. Then the authors use these patterns to detect causal relations from a domain-specific cor-

pus of corporate documents. The experiment on 32,545 documents in the Product Development-Customer Service (PD-CS) domain achieves 85% F-score, and outperforms the state-of-the-art model. Since implicit causality is more complicated than explicit causality, the system needs much more effort to define the three modes.

Table 4

Examples of causal patterns from Wikipedia in Ittoo and Bouma, 2013 [66]

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"

Inter-sentential Causality Khoo et al. [2] proposes a work of relied on four kinds of causal links and 2,082 causative verbs to construct a set of verbal linguistic patterns for CE. A computer program finds all parts of the document that match with any of the linguistic patterns, so it can identify causal relations within a sentence, as well as between adjacent sentences. However, because the lack of knowledge-based inferencing, the experiment on 1,082 sentences of Wall Street Journal newspapers only achieves 19% precision. Verbal-linguistic patterns are used to extract relations between mutations in viral genomes(cause) and HIV drugs(effect) in the system of [8]. Text retrieval phase sort out candidate intra- and inter-sentences if there are <mutation, relation, drug> triplets. Then text preprocessing phase simplify candidate sentences, and manually an-

alyze if there are causality by a list of pre-defined keywords, and finally relation extraction phrase apply eleven causality rules to form linguistic patterns. This system is used in five hospitals to find resistance data from medical literature. However, the complicated text preprocessing step, which includes replacing "known" terms, grouping mutation and drug names, normalizing sentences, severely limits the model's flexibility.

In Table 5, we summarize the above knowledge-based CE systems with primary information in detail, which include the causality forms, experiment datasets, linguistic cues, approaches, as well as the results.

5.2. Statistical machine learning-based approaches

Statistical machine learning-based approaches require less manual pre-defined patterns than knowledge-based approaches. They usually utilize third-party NLP tools (e.g., Stanford CoreNLP [67], Spacy [68], Stanza [69]) to generate a set of features for amounts of labelled data, and then use ML algorithms (e.g., support vector machine (SVM), maximum entropy (ME), naive bayes (NB), and logistic regression (LG)) to perform the classification. In the following paragraphs we explain how statistical machine learning techniques are used in CE systems.

Explicit Intra-sentential Causality Inspired by the previous work [70] in 2002, which utilize lexicon-syntactic patterns to infer causation in a semi-automatic way. One year later, Girju [16] proposes a model to detect causal relations in a QA system. This model focus on the most frequent explicit intra-sentential causality patterns, <NP1, verb, NPs>. where the verb is a simple causative, and then validate those patterns refer to causation through a set of features with decision tree (DT). Blanco et al. [19] only identify causality in the pattern of <*VerbPhrase, relator, Cause*>, where *relator* belongs to *because, since, after, as*. They use seven kinds of features with C4.5 DT on a popular question classification dataset, called as TREC, and achieve an F-score of 91%. Most errors in the system occur when the *relator* in the pattern is *as* or *after*. This means that the system only works well when the instances have very clear

Table 5

Knowledge-based approaches.

Systems	Causality Forms	Datasets	Linguistic cues	Techniques	Results
[14]	Explicit Intra-sentential	Technical texts in French	Causality connectors, surround noun phrases, ambiguity words	Pattern matches	Precision (85%)
[7]	Explicit Intra-sentential	130 abstracts from medical documents	Common causal links, domain-specific causative verbs	Pattern matches	Precision (68%)
[15]	Explicit Intra-sentential	150 years of news articles	Causality connector, syntactic constraint, POS tags	Rule matches	Precision (78%)
[66]	Implicit	thirty thousand documents in PD-CS domain	POS tags, syntactic parsed tree	Pattern matches	F-score (85%)
[2]	Inter-sentential	Four months of WSJ articles	Causal links, causative verbs, resultative constructions, conditional constructions, causative adverbs and adjectives	Patterns matches	Accuracy (68%)
[8]	Inter-sentential	PubMed abstracts	Nominal subject, passive subject, direct and indirect object, prepositional object, coordination and conjunction, adjectival complement	Rule matches	F-score (84%)

keywords, *because* or *since*. Kim et al. [64] combine probabilistic topic model with time-series causal analysis to mining causal topics. It iteratively refines topics, increasing the correlation between discovered topics with time series data. One experiment is to mining specific topics that are expected to affect the 2000 presidential election. Figure 2 shows several important issues (e.g., tax cuts, oil energy, and abortion). Such topics are also cited in political-related literature, which shows the efficiency of this model.

Fig. 2. Sample results in Kim et al. 2013 [64]

Implicit Causality In order to solve the challenge of lack of explicit keywords, the approach of [51] utilizes four kinds of features, which are production rules, dependency rules, word pairs and contextual, with a ME to recognize implicit discourse relations in PDTB. Experiment result indicates that the feature of production rules contribute the most to the implicit relation extraction, followed by word pairs, dependency rules, and contextual. But as the model tends to label many uncertain relations as causality, it leads to very low precision performance for the causal relation. Keskes et al. [24] design a ME model to learn causality in Arabic. Within the system, eight linguistic features make significant contributions to identifying implicit relations, like the modality feature to check if the sentence has Arabic modal words by a manually constructed lexicon. The experiment on newswire stories achieves the F-score of 80% and accuracy of 93%. However, these rich and complex feature lists rely heavily

on NLP tools, like Standard Arabic Morphological Analyzer, Stanford parser and various linguistic resources. Meanwhile, due to the characteristics of different languages, some well-extracted features may useless in other languages. Pechrisi and Kawtrakul [71] utilize verb-pair rules to train NB and SVM to mining implicit causality from Thai texts. WordNet and pre-defined plant disease information are used to collect the causative and effect verb concept set. Causative and effect verb concept in instance consist as the verb-pair rules. The experiment on 3,000 agricultural-related sentences obtain precision and recall of 86% and 70%, respectively. Unlike above two approaches that used rich set of features to represent the input instances, this model focus on collected background knowledge, but it cause the model only can be applied on a small amount of domain-specific texts.

Inter-sentential Causality March and Echihabi [72] utilize lexical pair probability to discriminate causality in inter-sentential forms. They use sentence connected keywords "Because of" and "Thus" to find candidate sentence pairs, and use pre-collected explicit causality nouns, verbs and adverbs to find the causal lexical pairs. Non-causal lexical pairs are obtained from randomly selected sentence pairs. Oh et al. [61] 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 to identify cue phrases for causality. Then, for each identified cue phrase, they extract three sentences as one causality candidate, which include 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.

With the availability of NLP tools, approaches to causality extraction based on statistical machine learning methods have become ubiquitous. Similar to Table 5, Table 6 lists the approaches discussed in the above paragraphs with their datasets, features or kernels, algorithms, and performances.

Table 6

Statistical machine learning based approaches.

System	Causality Form	Datasets	Features/Kernels	Algorithms	Results
[16]	Explicit Intra-sentential	TREC corpus	lexicon-syntactic patterns	DT	Precision (74%), Recall (89%)
[19]	Explicit Intra-sentential	TREC5	Relator, lexical clue,relator left and right modifiers, semantic class cause verb, cause/effect verb is potentially causal, semantic class effect verb, verb tense cause and effect verb	C4.5 DT	F-score (91%)
[51]	Implicit	PDTB	Contextual, constituent parse, dependency parse	ME	F-score (51%)
[24]	Implicit	90 documents from Arabic Treebank	Contexual, lexical, argument position, semantic relations, word polarity, named entities, anaphora and modality	ME	F-score (80%)
[71]	Implicit	1,000 examples from herb websites in Thai	causative-verb-phrase concept set, supporting causative verb set, effect-verb-phrase concept set	NB	Precision (86%), Recall (70%)
[72]	Inter-sentential	BLIPP	lexical pair probability,	NB	Accuracy (87%)
[61]	Inter-sentential	850 Japanese QA examples	Cue phrases, morphological, syntactic, c-marker features	CRF	F-score (77%)

5.3. Deep learning-based approaches

Neural Networks (NNs) are basic algorithm for deep learning (DL). Like a human's neural system, it is composed of neurons with the activation α in three kinds of layers: input, hidden, and output. Each neuron receives input from previous neurons and produces an output for later neurons. When an NN has multiple hidden layers, it is a 'deep' neural network, thus the process is referred to as 'deep learning' [73].

Compared with knowledge-based and statistical ML models, deep learning models map words and features into low-dimensional dense vectors, which can alleviate the feature sparsity problem. The most typical deep learning model include convolutional neural networks (CNNs), recurrent neural networks (RNNs) and long short-term momery (LSTM). Further more, use attention mechanism to selectively concentrating on relevant things, while ignoring others in deep learning model, makes deep learning models more effective. Later, unsupervised pre-training language models (PTMs), like BERT [74], which return contextualized embeddings for each token significantly improves performance on many NLP tasks [75]. Both CNNs and LSTMs can be viewed as sequential-based models, which embed semantic and syntactic information in local consecutive word sequences [76]. In comparison, graph-based models, like graph convolutional networks (GCNs) and graph attention networks (GATs), also capture researchers' attention.

Also, in the following paragraphs, we will introduce how deep learning architectures are used to solve the CE problem.

Explicit Intra-sentential Causality Xu et al. [49] utilize LSTM to learn relation representations along the words on the shortest dependency path (SDP) and obtained the F-score of 84% on the SemEval 2010 task 8 corpus. One year later, Wang et al. [50] propose a CNN with multi-level attention mechanisms to capture entity-specific and relation-specific information from the same dataset with [49]. Zhang et al. [47] propose a dependency tree-based GCN model to extract relationship in TACRED. To incorporate relevant information and remove irrelevant context, they apply a new tree pruning strategy that only

kept words directly connected to the shortest dependency path. The experiment achieves new state-of-the-art performance and shows the success of GCN on relation extraction task. However, the model is based on the way of NLP tools define the shortest dependency path, it will inevitably generate cascading errors. One year later, Kyriakakis et al. [17] incorporate ELMO and BERT, and use bidirectional GRU with self-attention (BIGRUATT) as a baseline, to detect causal sentences in four kinds of datasets. The experiment results show that context dependent PTMs improve the accuracy when the data contain hundreds of instances.

Implicit Causality Ponti and Korhonen [4] develop a feedforward neural network (FNN) model to extract causality. They combine positional and event-related features with the basic lexical feature as a enriched feature set. Positional features encode the distance between each word to the *cause* and *effect*, while event-related features accounts for the semantic of the input sentences. Experiment performance indicates that the enriched features, especially positional features, have positive impact on implicit causality identification. The authors of [18] believe the use of linguistic features may restrict to represent causality meaning, so they propose an LSTM model by only incorporating word embeddings as input. The experiment results indicate that their proposed model outperforms the state-of-the-art on PTDB corpus. As the first example in Figure 3, the verb *break* mostly has a causal meaning in training examples, but it is not a causative verb in this test sentence. Luckily, it is correctly classified by the proposed approach, while misclassified by B2, i.e., an SVM-based model.

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 Lan et al. 2017 [54]

Inter-sentential Causality Taking full advantage that LSTM can address the issue of learning long-range dependencies from a sequence of words, Man et al. [54], Guo et al. [55] and Jin et al. [77] use LSTM to capture long-dependency causality. Specifically, Jin et al. [77] use CNN to capture essential features from input examples and then utilize BiLSTM to obtain deeper contextual semantic information between cause and effect. Kruengkrai et al. [78] introduce a variant CNN, called multi-column CNN, to recognize event causalities. Based on the assumption that dependency path between cause and effect can be viewed as background knowledge, they use a wide range of dependency paths, no matter cause and effect appear within one sentence or appear in adjacent sentences, from web texts as extra input. Within this model, different columns represents different inputs, such as the event causality candidates, contextual information, and background knowledge. Each column has its independent convolutional and pooling layers. All the outputs are concatenated into a SoftMax function to perform the classification. The experiment result demonstrates that related background knowledge significantly improves the performance. To understand chemical-induced disease (CID) relations from biomedical articles, Sun et al. [10] propose two ME models to extract CIE at both intra- and inter-sentential levels, respectively. They construct training and test instances at inter-sentence level complied with three heuristic rules: (a) only two entities are not involved

in any intra-sentential instance are considered as inter-sentence level. (b) the sentence distance betwen two entities should be less than three. (c) keep the entities pair that in the nearest distance if there are multiple entities in one instance. Then they use ME classifier with various lexical features to extract this relationship from 1,500 medical articles.

Similarly, Table 7 lists the approaches discussed in the above paragraphs with their datasets, embeddings and/or features, algorithms, and performances.

5.4. Summary of systems with different causality forms

In this section, we reviewed how to use different techniques to extract different forms of causality. For extracting explicit causality, knowledge-based and statistical ML-based approaches utilize pre-defined clearly keywords or patterns. For example, Garcia et al. [14] use the connectives of *provoke, result, etc*, Khoo et al. [7] use common causal expressions for depression, and other approaches like [15, 16, 19] utilize different kinds of linguistic patterns. The focus of deep learning methods is how they design models to capture high-level features. While for extracting implicit causality, approaches need to construct more complicate patterns or features, like the three implicit patterns in [66], verb-pair features in [71], and positional features in [4]. However, no matter it is explicit or implicit causality, preparing these keywords, patterns or features is labor-intensive, and also severely restrict the portability and generalizability of CE to other resources. For intra- and inter-sentential causality, knowledge-based and statistical ML-based methods focus on how to prepare the training and test instances. Intra-sentential is more simple and straightforward that each instance is an independent sentence. In contrast, approaches for inter-sentential causality, like [2, 8, 72, 61, 10], need to utilize some specific words, patterns or cue phrases to find the candidate sentence pairs, even paragraphs.

6. Open problems and future directions

From the representative systems in Section 5, we can know that causal relation extraction has received increasing attention over the past decade. However,

Table 7

Deep Learning-based approaches.

System	Causality Form	Datasets	Embeddings/Features	Techniques	Results
[49]	Explicit Intra-sentential	SemEval 2010 Task 8	Word embeddings, POS, WordNet,grammar relation	LSTM	F-score (84%)
[50]	Explicit Intra-sentential	SemEval 2010 Task 8	Word embeddings	CNN	F-score (88%)
[47]	Explicit Intra-sentential	TACRED	Shortest dependency tree	GCN	F-score (68%)
[17]	Explicit Intra-sentential	SemEval-2010 Task 8, TimeBank, Event StoryLine, BioCausal	Word embeddings	BERT/ELMO with BI-GRUATT	F-score (92% and 93%) in SemEval-2010 Task 8, F-score (75% and 80% in TimeBank, F-score (73% and 75%) in Event StoryLine, F-score (86% and 87%) in BioCausal
[4]	Implicit	PDTB,CSTNews	Lexical, positional,event-related features	FNN	Accuracy (66%) in PDTB, Accuracy (80%) in CSTNews
[18]	Implicit	PDTB	Word embedding	LSTM	F-score (80%)
[78]	Inter-sentential	same with	Causal cue, contextual,	CNN	Precision (55%)
[10]	Inter-sentential	1,500 abstracts in medical	Contextual, dependency paths	CNN	Precision (52%)

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 from Dasgupta et al. [43], traditional ways to deal the above kinds of multiple causalities is dividing sentence into several sub-sentences that extracted causal pairs separately. This method is computationally expensive and cannot take into consideration the dependencies among causality pairs.

  The Tag2Triplet algorithm from [46] 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. Afterwards, it applies a Cartisian Product of the causal entities to generate possible causal triplets. In addition, [79, 80] 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 [81]. However, just as the size of the four benchmark datasets introduced in Section 3 is far from the size of a

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

satisfactory deep learning model, the annotated data in the real-world is very specific and small.

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. [82] 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 suffer from a large amount of noise labelled data. The survey of [23] 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 [83] propose a novel way for relation extraction from insufficient labelled 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. Afterwards, a random forest classifier is used to learn the relation types from these relation embeddings. This approach of combine unsupervised learning with supervised learning provide us another new idea of solve 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 [84]. All of these aspects need long-term research and exploration.

  Zeng et al. [85] 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 [86]. Similar approaches can be found in [87, 88].

7. 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 towards constructing a causality knowledge base. In this paper, we conducted a comprehensive review of CE in which we introduced four kinds of benchmark datasets and the evaluation metrics for this task. Afterwards, we reviewed existing approaches that use traditional or modern techniques to extract different causality forms. From Section 5, 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 corpura are quite different with the word distributions of specific domain corpora, the standard PTMs has been shown not to perform well in specialized domains [74]. In contrast, pre-training from scratch on domain-specific corpora, like BioBERT [89], BlueBERT [90], ClinicalBERT [91] and SciBERT [75], can alleviate this limitation. 2) The study of [47, 92] demonstrate the advantage of GCN in complex texts. Thus we can solve the CE problem by combining domain-specific PTMs with graph models.

References

[1] J. Cowie, W. Lehnert, Information extraction, Commun. ACM 39 (1) (1996) 80–91. doi:10.1145/234173.234209.

[2] C. S. G. KHOO, J. KORNFILT, R. N. ODDY, S. H. MYAENG, Automatic Extraction of Cause-Effect Information from Newspaper Text Without Knowledge-based Inferencing, Literary and Linguistic Computing 13 (4) (1998) 177–186.

[3] D.-S. Chang, K.-S. Choi, Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities, Inf. Process. Manage. 42 (3) (2006) 662–678. doi:10.1016/j.ipm.2005.04.004.

[4] E. M. Ponti, A. Korhonen, 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, 2017, pp. 25–30. doi:10.18653/v1/W17-0903.

[5] A. Balashankar, S. Chakraborty, S. Fraiberger, L. Subramanian, 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, 2019, pp. 2338–2348. doi:10.18653/v1/D19-1238.

[6] S. V. Cole, M. D. Royal, M. G. Valtorta, M. N. Huhns, J. B. Bowles, A lightweight tool for automatically extracting causal relationships from text, in: Proceedings of the IEEE SoutheastCon 2006, 2006, pp. 125–129. doi:10.1109/second.2006.1629336.

[7] C. S. G. Khoo, S. Chan, Y. Niu, 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, 2000, pp. 336–343. doi:10.3115/1075218.1075261.

[8] Q. C. Bui, B. o Nuallain, C. A. Boucher, P. M. Sloot, Extracting causal relations on hiv drug resistance from literature, BMC Bioinformatics 11 (1) (2010) 101–110.

[9] C. Mihaila, S. Ananiadou, Semi-supervised learning of causal relations in biomedical scientific discourse, Biomedical Engineering Online 13 (2) (2014) 1–24.

[10] L. Qian, G. Zhou, Chemical-induced disease relation extraction with various linguistic features, Database 2016 (2016) baw042. doi:10.1093/database/baw042.

[11] J. Qiu, L. Xu, J. Zhai, L. Luo, Extracting causal relations from emergency cases based on conditional random fields, Procedia Comput. Sci. 112 (C) (2017) 1623–1632. doi:10.1016/j.procs.2017.08.252.

[12] B. Beamer, A. Rozovskaya, R. Girju, Automatic semantic relation extraction with multiple boundary generation, in: Proceedings of the 23rd National Conference on Artificial Intelligence, AAAI Press, Chicago, Illinois, 2008, pp. 824–829.

[13] C. Khoo, S. Chan, Y. Niu, The many facets of the cause-effect relation, The Semantics of Relationships (2002) 51–70doi:10.1007/978-94-017-0073-3_4.

[14] D. Garcia, EDF-DER, IMA-TIEM, COATIS, an NLP system to locate expressions of actions connected by causality links, Vol. 1319, Springer, 2006, Ch. BFb0026799, pp. 347–352. doi:10.1007/BFb0026799.

[15] K. Radinsky, S. Davidovich, S. Markovitch, Learning causality for news events prediction, WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web (2012) 909–918doi:10.1145/2187836.2187958.

[16] R. Girju, Automatic detection of causal relations for question answering, in: Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering, Vol. 12 of MultiSumQA '03, Association for

Computational Linguistics, USA, 2003, p. 76–83. doi:10.3115/1119312.1119322.

[17] M. Kyriakakis, I. Androutsopoulos, A. Saudabayev, J. Ginés i Ametllé, Transfer learning for causal sentence detection, in: Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, Florence, Italy, 2019, pp. 292–297. doi:10.18653/v1/W19-5031.

[18] E. Martínez-Cámara, V. Shwartz, I. Gurevych, I. Dagan, Neural disambiguation of causal lexical markers based on context, in: IWCS 2017 — 12th International Conference on Computational Semantics — Short papers, 2017.

[19] E. Blanco, N. Castell, D. Moldovan, Causal relation extraction, in: Proceedings of the International Conference on Language Resources and Evaluation, Marrakech, Morocco, 2008, pp. 310–313.

[20] Q. Zhang, M. Chen, L. Liu, A review on entity relation extraction, in: 2017 Second International Conference on Mechanical, Control and Computer Engineering (ICMCCE), Vol. 1, 2017, pp. 178–183. doi:10.1109/ICMCCE.2017.14.

[21] R. A. Kadir, B. Bokharaeian, Overview of biomedical relations extraction using hybrid rule-based approaches, Journal of Industrial and Intelligent Information 1 (3) (2013) 169–173. doi:10.12720/jiii.1.3.169-173.

[22] D. Zhou, D. Zhong, Biomedical relation extraction: From binary to complex, Computational and mathematical methods in medicine 24 (2014) 298–473. doi:10.1155/2014/298473.

[23] A. Smirnova, P. Cudré-Mauroux, Relation extraction using distant supervision: A survey, ACM Comput. Surv. 51 (5) (2018) p. 35. doi:10.1145/3241741.

[24] I. Keskes, F. B. Zitoune, L. Belguith, Learning explicit and implicit arabic discourse relations, Journal of King Saud University - Computer and Information Sciences archive 26 (2014) 398–416.

[25] R. Wu, Y. Yao, X. Han, R. Xie, Z. Liu, F. Lin, L. Lin, M. Sun, 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, 2019, pp. 219–228. doi:10.18653/v1/D19-1021.

[26] N. Asghar, Automatic extraction of causal relations from natural language texts: A comprehensive survey, CoRR abs/1605.07895.

[27] B. Rink, C. Bejan, S. Harabagiu, Learning textual graph patterns to detect causal event relations, in: Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference, 2010, pp. 265–270.

[28] A. Sorgente, G. Vettigli, F. Mele, Automatic extraction of cause-effect relations in natural language text, DART@ AI* IA 1109 (2013) 37–48.

[29] X. Yang, K. Mao, Multi level causal relation identification using extended features, Expert Syst. Appl. 41 (16) (2014) 7171–7181. doi:10.1016/j.eswa.2014.05.044.

[30] S. Bethard, J. H. Martin, 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, 2008, pp. 177–180.

[31] B. Barik, E. Marsi, P. Ozturk, Event causality extraction from natural science literature, Research in Computing Science 117. doi:10.13053/rcs-117-1-8.

[32] R. Girju, P. Nakov, V. Nastase, S. Szpakowicz, P. Turney, D. Yuret, Semeval-2007 task 04: Classification of semantic relations between nominals, in: Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval '07, Association for Computational Linguistics, USA, 2007, p. 13–18.

[33] I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, S. Szpakowicz, 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, 2010, pp. 33–38.

[34] R. Prasad, E. Miltsakaki, N. Dinesh, A. Lee, A. Joshi, The penn discourse treebank 2.0 annotation manual, IRCS technical reports series 203. Philadelphia: University of Pennsylvania ScholarlyCommons. (2007) 105.

[35] S. Liang, W. Zuo, Z. Shi, S. Wang, A multi-level neural network for implicit causality detection in web texts, CoRR abs/1908.07822. arXiv:1908.07822.

[36] Y. Zhang, V. Zhong, D. Chen, G. Angeli, C. D. Manning, 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, 2017, pp. 35–45. doi:10.18653/v1/D17-1004.

[37] S. Pyysalo, F. Ginter, J. Heimonen, J. Björne, J. Boberg, J. Järvinen, T. Salakoski, Bioinfer: A corpus for information extraction in the biomedical domain, BMC bioinformatics 8 (2007) 50. doi:10.1186/1471-2105-8-50.

[38] H. Gurulingappa, A. Mateen, A. Roberts, J. Fluck, M. Hofmann-Apitius, L. Toldo, Development of a benchmark corpus to support the automatic

extraction of drug-related adverse effects from medical case reports, Journal of Biomedical Informatics http://dx.doi.org/10.1016/j.jbi.2012.04.008. doi:10.1016/j.jbi.2012.04.008.

[39] F. Li, M. Zhang, G. Fu, D. Ji, A neural joint model for entity and relation extraction from biomedical text, BMC Bioinformatics 18. doi:10.1186/s12859-017-1609-9.

[40] F. Li, Y. Zhang, M. Zhang, D. Ji, Joint models for extracting adverse drug events from biomedical text, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16, AAAI Press, 2016, p. 2838–2844.

[41] F. Li, M. Zhang, G. Fu, D. Ji, A neural joint model for entity and relation extraction from biomedical text, BMC Bioinformatics 18.

[42] R. Girju, P. Nakov, V. Nastase, S. Szpakowicz, P. Turney, D. Yuret, Classification of semantic relations between nominals, Language Resources and Evaluation 43 (2) (2009) 105–121.

[43] T. Dasgupta, R. Saha, L. Dey, A. Naskar, 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, 2018, pp. 306–316.

[44] C. Alt, M. Hübner, L. Hennig, Improving relation extraction by pre-trained language representations, in: Automated Knowledge Base Construction (AKBC), 2019, p. 18.

[45] S. Zhao, T. Liu, S. Zhao, Y. Chen, J.-Y. Nie, Event causality extraction based on connectives analysis, Neurocomputing 173 (2016) 1943–1950.

[46] Z. Li, Q. Li, X. Zou, J. Ren, Causality extraction based on self-attentive bilstm-crf with transferred embeddings, Neurocomputing 423 (2021) 207 – 219. doi:https://doi.org/10.1016/j.neucom.2020.08.078.

[47] Y. Zhang, P. Qi, C. D. Manning, 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, 2018, pp. 2205–2215. doi:10.18653/v1/D18-1244.

[48] P. Pakray, A. Gelbukh, An open domain causal relation detection from paired nominal, in: 13th Mexican International Conference on Artificial Intelligence (MICAI-2014), Vol. 8857, Nature-Inspired Computation and Machine Learning, Chiapas, Mexico, 2014, pp. 261–271. doi:10.1007/978-3-319-13650-9-24.

[49] Y. Xu, L. Mou, G. Li, Y. Chen, H. Peng, Z. Jin, 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, 2015, pp. 1785–1794. doi:10.18653/v1/D15-1206.

[50] L. Wang, Z. Cao, G. de Melo, Z. Liu, 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, 2016, pp. 1298–1307. doi:10.18653/v1/P16-1123.

[51] Z. Lin, M.-Y. Kan, H. T. Ng, 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, 2009, pp. 343–351.

[52] C. Hidey, K. McKeown, 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, 2016, pp. 1424–1433. doi:10.18653/v1/P16-1135.

[53] J. Chen, Q. Zhang, P. Liu, X. Qiu, X. Huang, 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, 2016, pp. 1726–1735. doi:10.18653/v1/P16-1163.

[54] M. Lan, J. Wang, Y. Wu, Z.-Y. Niu, H. Wang, 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, 2017, pp. 1299–1308. doi:10.18653/v1/D17-1134.

[55] F. Guo, R. He, J. Dang, Implicit discourse relation recognition via a bilstm-cnn architecture with dynamic chunk-based max pooling, IEEE Access 7 (2019) 169281–169292.

[56] S. Pyysalo, F. Ginter, V. Laippala, K. Haverinen, J. Heimonen, T. Salakoski, On the unification of syntactic annotations under the Stanford dependency scheme: A case study on BioInfer and GENIA, in: Biological, translational, and clinical language processing, Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 25–32.

[57] D. Chen, Y. Li, K. Lei, Y. Shen, Relabel the noise: Joint extraction of entities and relations via cooperative multiagents, ArXiv abs/2004.09930.

[58] N. Kang, B. Singh, Q.-C. Bui, Z. Afzal, E. M. van Mulligen, J. Kors, Knowledge-based extraction of adverse drug events from biomedical text, BMC bioinformatics 15 (2014) 64. doi:10.1186/1471-2105-15-64.

[59] S. Zhao, M. Hu, Z. Cai, F. Liu, Modeling dense cross-modal interactions for joint entity-relation extraction, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, International Joint Conferences on Artificial Intelligence Organization, 2020, pp. 4032–4038, main track. doi:10.24963/ijcai.2020/558.

[60] J. Wang, W. Lu, 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, 2020, pp. 1706–1721. doi:10.18653/v1/2020.emnlp-main.133.

[61] J.-H. Oh, K. Torisawa, C. Hashimoto, M. Sano, S. De Saeger, K. Ohtake, 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, Vol. 1, Sofia, Bulgaria, 2013, pp. 1733–1743.

[62] J.-H. Oh, K. Torisawa, C. Hashimoto, R. Iida, M. Tanaka, J. Kloetzer, A semi-supervised learning approach to why-question answering, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16, AAAI Press, 2016, p. 3022–3029.

[63] J.-H. Oh, K. Torisawa, C. Kruengkrai, R. Iida, J. Kloetzer, 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, 2017, pp. 415–424. doi:10.1145/3018661.3018737.

[64] H. Kim, M. Castellanos, M. Hsu, C. Zhai, T. Rietz, D. Diermeier, 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, 2013, pp. 885–890. doi:10.1145/2505515.2505612.

[65] C. W. J. Granger, Investigating causal relations by econometric models and cross-spectral methods, Econometrica 37 (3) (1969) 424–438.

[66] A. Ittoo, G. Bouma, Minimally-supervised learning of domain-specific causal relations using an open-domain corpus as knowledge base, Data &

Knowledge Engineering 88 (2013) 142–163. doi:10.1016/j.datak.2013.08.004.

[67] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, D. McClosky, 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, 2014, pp. 55–60. doi:10.3115/v1/P14-5010.

[68] M. Honnibal, I. Montani, spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing, To appear.

[69] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, 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, 2020, pp. 101–108. doi:10.18653/v1/2020.acl-demos.14.

[70] A. Sobrino, J. A. Olivas, C. Puente, Mining answers for causal questions in a medical example, in: 2011 11th International Conference on Intelligent Systems Design and Applications, 2011, pp. 432–437. doi:10.1109/ISDA.2011.6121694.

[71] C. Pechsiri, A. Kawtrakul, R. Piriyakul, Mining causality knowledge from textual data, in: Proceedings of the 24th IASTED International Conference on Artificial Intelligence and Applications, AIA'06, ACTA Press, USA, 2006, p. 85–90.

[72] D. Marcu, A. Echihabi, 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, 2002, pp. 368–375. doi:10.3115/1073083.1073145.

[73] K. O'Shea, R. Nash, An introduction to convolutional neural networks, ArXiv e-printsarXiv:1511.08458.

[74] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT, 2019, pp. 4171–4186.

[75] I. Beltagy, K. Lo, A. Cohan, 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, 2019, pp. 3615–3620.

[76] L. Yao, C. Mao, Y. Luo, Graph convolutional networks for text classification, in: In 33rd AAAI Conference on Artificial Intelligence, 2019, pp. 7370–7377.

[77] X. Jin, X. Wang, X. Luo, S. Huang, S. Gu, Inter-sentence and implicit causality extraction from chinese corpus, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Vol. 12084, Springer, 2020, pp. 739–751. doi:10.1007/978-3-030-47426-3-57.

[78] C. Kruengkrai, K. Torisawa, C. Hashimoto, J. Kloetzer, J.-H. Oh, M. Tanaka, 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, 2017, p. 3466–3473.

[79] F. Christopoulou, M. Miwa, S. Ananiadou, 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, 2018, pp. 81–88. doi:10.18653/v1/P18-2014.

[80] H. Wang, M. Tan, M. Yu, S. Chang, D. Wang, K. Xu, X. Guo, S. Potdar, 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, 2019, pp. 1371–1377. doi:10.18653/v1/P19-1132.

[81] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.

[82] M. Mintz, S. Bills, R. Snow, D. Jurafsky, 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, 2009, pp. 1003–1011.

[83] H. Huang, R. Wong, Deep embedding for relation extraction on insufficient labelled data, in: 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1–8. doi:10.1109/IJCNN48605.2020.9207554.

[84] F. Christopoulou, M. Miwa, S. Ananiadou, 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, 2019, pp. 4927–4938.

[85] S. Zeng, R. Xu, B. Chang, L. Li, 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, 2020, pp. 1630–1640.

[86] Y. Yao, D. Ye, P. Li, X. Han, Y. Lin, Z. Liu, Z. Liu, L. Huang, J. Zhou, M. Sun, 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, 2019, pp. 764–777. doi:10.18653/v1/P19-1074.

[87] H. Minh Tran, M. T. Nguyen, T. H. Nguyen, 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, 2020, pp. 4561–4567.

[88] D. Wang, W. Hu, E. Cao, W. Sun, 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, 2020, pp. 3711–3721.

[89] L. Jinhyuk, Y. Wonjin, Kim, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2019) 1234–1240.

[90] Peng, Yifan, Yan, Shankai, 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, 2019, pp. 58–65.

[91] K. Huang, J. Altosaar, R. Ranganath, Clinicalbert: Modeling clinical notes and predicting hospital readmission, arXiv:1904.05342.

[92] Z. Guo, Y. Zhang, W. Lu, 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, 2019, pp. 241–251. doi:10.18653/v1/P19-1024.