A Review of dataset and labeling methods for causality extraction

Jinghang Xu[1,2], **Wanli Zuo**[1,2], **Shining Liang**[1,2], **Xianglin Zuo**[1,2*]
[1]College of Computer Science and Technology, Jilin University
[2]Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education
(xujh17, liangsn19)@mails.jlu.edu.cn
zuowl@jlu.edu.cn, 295228473@qq.com

Abstract

Causality represents the most important kind of correlation between events. Extracting causality from text has become a promising hot topic in NLP. However, there is no mature research systems, evaluation rules and datasets for public evaluation. Moreover, there is a lack of unified causal sequence labeling methods, which constitute the key factors that hinder the progress of causality extraction research. We survey the limitations and shortcomings of existing causality research field comprehensively from the aspects of basic concepts, extraction methods, experimental data, and labeling methods, so as to provide reference for future research on causality extraction. We summarize the existing causality datasets, explore their practicability and extensibility from multiple perspectives. Aiming at the problem of causal sequence labeling, we analyze the existing methods of causal sequence labeling, with a summarizations of its regulation. Multiple candidate causal labeling sequences are put forward according to labeling controversy to explore the optimal labeling method through experiments, and suggestions are provided for selecting labeling method.

1 Introduction

Causality represents a kind of relationships between "cause" and "effect", stating that the happening of causes will trigger the happening of effects, which constitutes the basics for inference and reasoning. In early days, people found causal relations from production process and daily lives manually, slow and inaccurate. The ever growing web resources have made it feasible to automatically extracts causality from text, triggering an emerging and hot topic in NLP, with abundant downstream application tasks such as event detection and prediction (Radinsky et al., 2012), questions answering (Hashimoto et al., 2014). For example, the new coronary pneumonia has drawn global wide attention recently, the causal relationship between "new coronavirus" and "eating wild animals" can be inferred by an expert system with causal knowledge, and it may be predicted that "new coronary pneumonia" can lead to "death". Artificial intelligence can be used to assist medical research and strengthen prevention and treatment during early stage of infection. Therefore, causal relation extraction is a basic task in text mining, which is driven by human's instinctive desire for knowledge.

Many researchers have devoted themselves to the research of causal relationship extraction, however, it is still a new field with some open problems to slove and publicly evaluated datasets. The existing relevant studies each has its own research system that cannot be compared horizontally, so the systematic research system is the key factors for the progress of causality extraction research. Targeting at the limitations and problems in causality research, we comprehensively review the basic concepts, extraction methods, experimental data, labeling methods from multiple perspectives, for future research on causality extraction as a reference. The main contributions of this paper are listed as follows:

- We summarize the concept of causality and the existing research methods of causality extraction.

*Corresponding author.

Proceedings of the 28th International Conference on Computational Linguistics, pages 1519–1531
Barcelona, Spain (Online), December 8-13, 2020

- Regarding the defects and shortcomings of the existing experimental data for causality, we summarize the publicly available dataset, analyzing them from multiple aspects.
- Targeting at the causal research method of sequence labeling, we comprehensively summarize and analyze the existing methods. Multiple candidate causal labeling sequences are set to avoid the ambiguity of labeling, and the optimal method is explored through experiments. Our results show that the "core word" causal sequences labeling method archives the best effect, giving suggestions for the selection of labeling method.

2 Causality and causality extraction

2.1 Causality

Causality is the correspondence between "cause" and "effect", the "cause" is the producer of "effect" and the "effect" is the outcome of "cause". There are many elements to express causal semantics.

2.1.1 Causal unit

Different studies have different demarcations of causal boundary, and different text sentences are also applicable to different "causal unit". We summarize four kinds of causal units:

- **Word**: In the sentence texts which word can fully express causal semantics, word can be used as the unit for causal boundary division. Such as sample "$\langle e1\rangle$Suicide$\langle/e1\rangle$ is one of the leading causes of $\langle e2\rangle$death$\langle/e2\rangle$.", the word "suicide" is the cause, the word "death" is the effect.
- **Phrase**: In some sentences, the causal semantics of "word" is not complete, so it need to take "phrase" as the unit. Such as sample "$\langle e1\rangle$Financial stress$\langle/e1\rangle$ is one of the main causes of $\langle e2\rangle$divorce$\langle/e2\rangle$.", the phrase "financial stress" is a more complete way to express causal semantics than the word "stress".
- **Clause**:In some sentences that cannot extract the core word or phrase for causality, the "clause" should be used as the unit. Such as sample "$\langle e1\rangle$We play with a steady beat$\langle/e1\rangle$ so that $\langle e2\rangle$dancers can follow it$\langle/e2\rangle$.", it is impossible to extract the core word or phrase from the text to express causality.
- **Event**: An event is defined as a fact that takes place at a particular time and place, with several actors and the performance of several action characteristics. Such as sample "$\langle e1\rangle$A car travelling from Guizhou to Guangdong collided head-on with a bus$\langle/e1\rangle$ results the $\langle e2\rangle$ten people, six men and four women, including the driver, died at the scene$\langle/e2\rangle$.", semantically, the event 1 causes the event 2.

2.1.2 Causal connectives

Form	Explicit Connectives	Ambiguous Connectives
Verb	cause, result, arise, trigger	increase, affect, effect, make, induce, derive, reveal
Conjunction	because, so	hence, therefore, thus, thereby, since
Preposition	for, because of,	from, as, with, through, after
Adverb	accordingly, consequently	
Verb Phrase	result in(from), lead to	stem from, bring about, give rise to, thanks to
Prepositional Phrase	for the reason(alone), as a result of,	owing to, due to, in consequence of, in view of, as a consequence of, on account of, in as much as
Clause	that's why, the result is, by reason that, so that	on this(that) account, in this way, in that, now that, on the grounds that, for fear that, if... then

Table 1: Summary of common causal connectives in English

2.1.3 Special causal concept

- Sufficient causality (Luo et al., 2016): The "cause" sufficiently leads to the "effect". Such as the causal pair (storm, damage), the occurrence of "snowstorm" will inevitably lead to "damage", but "damage" is not necessarily caused by "snowstorm", but may be "earthquake", "flood" and so on.

- Necessary causality (Luo et al., 2016): The occurrence of "effect" inevitably leads to the embodiment of "cause". Such as the causal pair (rainfall, flooding), the occurrence of "flood", the high probability is the cause of "rain", but "rain" does not necessarily lead to "flood", but also can lead to "traffic jam".

- Temporal causality and Granger causality (Granger, 1988): In addition to "cause" and "effect", there is also "time" factor in the causality, Granger causality means the cause must precede the effect.

2.2 Complex causal correspondence

2.2.1 Multiple Causality

According to the correspondence between cause and effect, causality can be divided as one-cause and one-effect, one-cause and multi-effect, multi-cause and one-effect, multi-cause and multi-effect. With the increasing number of causality entities, the causal relationship shows special forms.

Embedded-causality A particular kind of multiple causality that manifests itself in the causal semantic (Li et al., 2019). There are entities with different causal semantic in different causal pairs. Such as sample "He testified that cause of $\langle e1\rangle$death$\langle/e1\rangle$ was $\langle e2\rangle$massive bleeding$\langle/e2\rangle$ into the blood sac of the heart caused by $\langle e3\rangle$a stab wound$\langle/e3\rangle$ on the left chest", the two causal pairs "(e2,e1), (e3,e2)" mean that "massive bleeding" is the cause of "death" and the effect of "a stab wound". There is a causal chain "a stab wound→massive bleeding→death" in the causality, so we also call it "chain-causality".

Cross-causality A particular kind of multiple causality that manifests itself in the causal position. In traditional causality, causes and effects appear continuously in the text, which means the adjacent two causal entities are assumed to be the same causal pair. However, in some special multiple causality, a causal pair appears intermittently, and multiple causal pairs "cross" in the text, we call it as "cross-causality". Such as sample "A $\langle e1\rangle$fire$\langle/e1\rangle$ broke out in the school, due to the help of $\langle e2\rangle$wind$\langle/e2\rangle$, the fire department $\langle e3\rangle$dispatched$\langle/e3\rangle$ 8 fire engines and 33 commanders to $\langle e4\rangle$help$\langle/e4\rangle$" with two causal pairs "(e1,e3), (e2,e4)", between the two entities "fire" and "dispatched" in the first causal pair, the cause "wind" in the second set appears. If only the label sequence "CCEE" is used for labeling, the corresponding causality cannot be recognized, so it needs to be identified through other labeling methods.

2.2.2 Intra-sentence, across-sentence, across-paragraph causality

Causality can be classified according to whether contain causal connectives (Ittoo and Bouma, 2011).

Explicit causality (1) with explicit connective: with the verbs such as "because" which have obvious causal meanings. (2) with ambiguous connectives: there are three forms such as "increase" (verbs with causal meaning can be realized by means of instrumental verb patterns), "generate(by)" (making causal agency inseparable from a situation), and "due to" (non-verb mode).

Implicit causality Implicit causality is often expressed in absence of causal connectives in the text, or there is only one entity of the cause or effect, and the other hide in the semantics of context.

2.3 Causality extraction

Many scholars have devoted themselves to the causality extraction researches, however, existing approaches are very scattered with no systematic research system. The early research used pattern matching to extract causality according to the structural characteristics of causal text. With the advancement of machine learning theory, the scope of text forms keeps expanding to multiple forms. With the increasing popularity of deep learning, CNN (convolutional neural network), RNN (recurrent neural network) and other deep learning models have been added to the research on causality extraction.

Based on the existing scattered research results, we summarize and sort out the three causal research fields: text classification, relation extraction, sequence labeling.

Text classification method Text classification is used to automatically classify a given text according to a certain classification system. Causality extraction based on text classification method is to classify text sentences according to whether they contain causal relation (Blanco et al., 2008; Hidey and McKeown, 2016; Kayesh et al., 2019; Paul, 2017). The method does not need to extract the causal events or entities in the text, only judge whether the whole sentence contains the causality, which is applicable to the causal data which is difficult to extract the events or entities from the text (the causal unit is clauses).

Relation extraction method Relation extraction is to judge whether the entity pair in a sentence has the specified relation, which is a dichotomy problem. Causality extraction based on relation extraction method is determining whether the causal pair given in the text has a causal relationship, which is applicable to the sentences in which the causal entity has been extracted. The existing models have naive Bayesian mode (Zhao et al., 2016), BGRU (bidirectional gated recurrent network) (Feng et al., 2018), multicolumn CNN (Kruengkrai et al., 2017), K-CNN (Knowledge-oriented CNN) (Li and Mao, 2019).

Sequence labeling method For a one-dimensional linear input sequence, each element in the linear sequence is labeled with a tag in the tag set. If the tag is set as causal semantic tag, the problem of causality extraction is transformed into sequence labeling, which is to label the causal tag for each word in the sentence, extracting the causality entity and determining the direction of causal relation. The existing models have CRFs (Fu et al., 2011), SCIFI (Self-Attentive Bi-LSTM-CRF with Flair Embeddings) (Li et al., 2019), L-BL (linguistically informed Bi-LSTM) (Dasgupta et al., 2018) and Bi-LSTM+CRF+ S-GAT (graph attention networks base on syntactic dependency graph) (Xu et al., 2020).

Except the above three common causality extraction methods, there are many new research methods.

- **Pattern matching**: Extract the causality with explicit mark(Garcia, 1997; Khoo et al.,1998; Girju, 2003).
- **Temporal causality**: With the "time" feature (Hashimoto et al.2012; Qiu et al.2017; Kang et al.2017).
- **Causal relationship of event**: Some studies for causality are conducted as the unit of "event". (Fu et al.,2011; Hashimoto et al.2012; Do et al.2011; Zhao et al.2017; Mirza and Tonelli 2014; Mirza 2014)
- **Causal network**: The network created by the causality events or entities extracted from a large number of corpus texts to improve the accuracy of the model or to be used for specific research. (Kayesh et al., 2019; Paul, 2017; Osoba and Kosko, 2019; Kocaoglu et al., 2017; Nordon et al., 2010; Yeo et al.,2018)
- **Multiple languages causality**: In addition to English, there have also been studies on causality extraction in other languages, including Chinese (Fu et al.,2011), Japanese (Hashimoto et al., 2012; Hashimoto et al., 2014; Hashimoto et al., 2015), German (Tina et al.2014), Arabic (Sadek and Meziane, 2016) ,etc.

Causal relationship extraction has been widely used in various fields, such as the biomedical information (Nordon et al., 2010; Khoo et al., 2000; Raja et al., 2013; Fluck et al., 2016; Casillas et al., 2016; Bollegala et al., 2018), psychology (White, 1990), analysis of social discrimination (Qureshi et al., 2016), log query (Sun et al., 2007), image (Kocaoglu et al., 2017), etc. Therefore, causal relationship extraction, as an important task, has been infiltrated into the research of various fields.

3 Dataset for causality

Experimental data is the basis of all research, however, there is currently no publicly evaluated dataset. which is one of the key factors that hinder the progress of causality research. According to the deficiencies for the causal dataset construction, we analyze the existing dataset for the subsequent studies.

3.1 Existing public dataset

SemEval (Semantic Evaluation) A public dataset for relation extraction, which has multiple relationships including instrument-agency, product-producer, etc., the cause-effect is the subject of this paper. There are 1368 sentences with causality and 107 sentences without causality. The advantages are: (1) Strong credibility. (2) Clarify "cause" and "effect": The causal and effect can be directly obtained according to the marked causal entities and the direction of causality. (3) Wide range of application: Simple data processing can be employed in the traditional three methods of causal extraction. While it has many disadvantages: (1) Small data amount : It cannot meet the needs for experiment, researches have extended it (Feng et al.2018; Li and Mao 2019; Xu et al., 2020). (2) Sample imbalance: The ratio of positive and negative cases of causality is about 12:1, other relationships should be added as the negative cases in the classification method (Feng et al.2018; Silva et al.2017). (3) In-consistent labeling standards: There is no unified labeling standard, so most related works have their own labeling methods (Dasgupta et al., 2018; Li et al., 2019; Xu et al., 2020). (4) It is a dataset for relation extraction, which only focus on whether the entities in the label have the relationship, but ignores the entities outside the label, which needs to be expanded manually (Li et al., 2019; Xu et al., 2020).

Causal-TimeBank and Event StoryLine (Li and Mao 2019) The format and usage method are basically the same as SemEval. However, the author has only disclosed a small amount of his data.

Altlex (Hidey and McKeown, 2016) It extracted the sentence text with causality from the English wikipedia corpus. There are 4,595 causal sentences and 39,645 no-causal sentences, with large data amount. However, it can only be applied to text classification method according to the unmarked entities.

CEC (Chinese event causality)(Fu et al., 2011) Collecting five types of emergencies from the Internet as raw materials, artificially label corpus events, with 200 articles and 340 sets of causal relationship. The advantages of this dataset are: (1) The only publicly available Chinese causal dataset. (2) Show the text in detail in the form of xml tags, so as to provide reference for the processing of experimental data. (3) With abundant types of causality, covering the widest range of causalities so far. At the same time, it alse had some disadvantages: (1) Small amount. (2) Require complex preprocessing to apply to traditional methods.

SCIFI (Li et al., 2019) In view of the defects and shortcomings of SemEval, SCIFI extended one causality to multiple causality, word to phrase. There are 1270 and 3966 sentences with and without causality respectively in the dataset (SCIFI is the name of the model, we call the dataset as SCIFI).

3.2 Summary of existing dataset for research methods

We summarize the application of existing six publicly available causality datasets in the three traditional researches is shown in Table 3.2. If the sequence labeling method is adopted, all sentences with causality in the dataset should be taken as the experimental data. For the two classification methods, all sentences with causality can be taken as positive cases, and sentences without causality and with other relationships can be taken as negative cases for experiment. If the causal entity is marked in the text, remove the entity labels leaving the pure sentence to judge whether it has causality (text classification), determine whether the given causal pairs has the causality directly (relation extraction), and set the label rules to generate the causal tag sequence (sequence labelling). If there is no mark of causal entity in the dataset, which can only adopt the text classification method to extract causal relation.

4 Causal sequence labeling method

For the causality extraction based on sequence labeling method, how to label the causal sequence is the key to the formation of experimental data. However, there is no fixed labeling rule, and horizontal comparison is impossible at present. We summarize the existing labeling methods and conduct multi-angle analysis, as a reference for the follow-up research.

Dataset	Text Classification	Relation Extraction	Sequence Labeling
SemEval	Remove the entity mark	✓	Label process
Causal-TimeBank	Remove the entity mark	✓	Label process
altlex	✓	×	×
CEC	Complex process	Complex process	Label process
SCIFI	Remove the entity mark	✓	Label process
ESC	×	×	✓

Table 2: Summary for the application of research methods for existing causal datasets. Which "✓"("×") means that it can(cannot) be directly applied, and other words mean the processing method

4.1 Existing causal sequence labeling methods

Phrase boundary (Li et al., 2019) The method uses the "BIO" (Begin, Inside, Other) entity boundary tag in NER to mark the causal phrase boundary. Three causality semantic tags "C"(Cause), "E"(Effect), and "Emb(Embedded-causality) are used to represent the causal semantics, , the "Emb" tag is introduced to solve the problem that embedded-causality cannot give accurate labels.

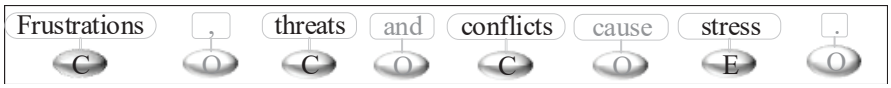

Figure 1: Example for the causal labeling method of phrase boundary

Clause subscript (Dasgupta et al., 2018) "C"(cause), "E"(effect), "CC" (causal connectives) and "N"(None) are used for causal semantics. The boundary is divided by clauses and different causal pairs are distinguished by the subscript in the form of "id".

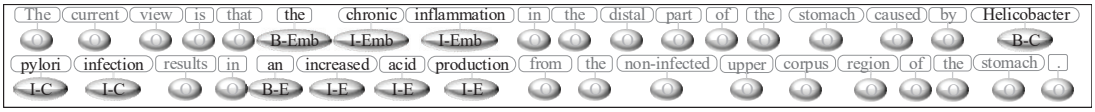

Figure 2: Example for the causal labeling method of clause subscript. The tags "C1", "E1" and "CC1" belong to the first set of causal pair, and the compound tag "E1C2" represents the entity is the effect of the first causal pair, and the cause of the second causal pair, deftly solving the embedded-causality

Advantages: (1) Using clause as the unit solves the problem that the text cannot extract the core word or phrase for causality. (2) A wide range of causal categories is extracted by means of the subscript.

Disadvantages: (1) In terms of clause unit, lots of irrelevant words are introduced in causal text. The sentence is only cut into two parts according to the causal conjunction in most text, the boundary is too loose and the essence of causal extraction is lost. (2) Lots of new causal tags are introduced due to the subscripts, the tags of sentences are distributed unevenly, which makes it difficult to train the feature.

Event sequence block (Fu et al., 2011) With tags "C", "E" and "N" to represent "Cause", "Effect" and "None", takes the event indicator as the representative of the overall event, without labeling other texts in the event. The "BIO" tag is used to distinguish the boundary of causal event blocks which divides the different causal pairs, solving the multiple causality problem. Core word (Xu et al., 2020): The method uses "C"(cause), "E"(effect) and "O" (Other) for causal tag, and core word for labeling unit. It solves multiple causality according to the number of tags "C", "E" and the embedded-causality by the means of "make rules to specify existing tag"

4.2 Three steps of labeling rules

According to the existing labeling methods, we summarize the labeling rules for the three steps. By setting the causality semantic tags, we can determine whether the entity is a cause or an effect, and then divide the entity boundary range according to the causality labeling unit. Finally, we can deal with some special causality by setting the label, unit division and introducing other methods.

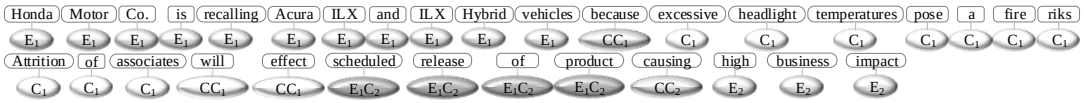

Figure 3: Example for the causal labeling method of event sequence block (translate it into English)

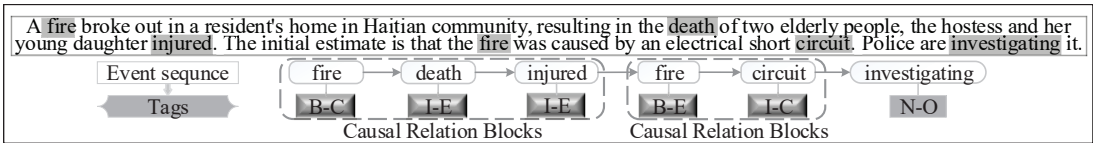

Figure 4: Example for the causal labeling method of core word. The sentence is three-cause and one-effect

Setting the causal semantic label In different studies, the causality semantic tag is roughly same. The first letters "C" and "E" of the words "cause" and "effect", the abbreviations "Emb", "CC" of "Embedded-Causality", "causal connectives" and other causal words are as the causal semantic tags.

Establishing the causal labeling units There are three causal labeling units: core word (Xu et al., 2020), phrase (Dasgupta et al., 2018) and clause (Li et al., 2019). There are differences in the labeling sequence and completeness of causal semantics according to the different causal labeling units.

Label Units	Label boundary strictness	Introducing no causality semantics	Causal semantics completeness	Label controversial
Word	Heavy	Slight	Incomplete	Moderate
Phrase	Moderate	Moderate	Moderate	Heavy
Clause	Slight	Heavy	Complete	Slight

Table 3: Causal labeling unit summary

Special causal treatment Through setting causal tags, units and introducing other methods, the existing four labeling methods can solve the labeling problems of some special causality.

4.3 Summary and analysis of the existing labeling methods

We comprehensively summarized and analyze existing four labeling methods from multiple angles in the Table 4.3. The more tag complex is, the clearer of tag semantics and the wider scope of causality is. The stricter for labeling boundary is, the less complete of causal semantic expression is, and less no-causal semantic is. The core word method sacrifices the clarity and completeness of causality for the simplicity and purity of labeling. On the contrary, the clause subscript uses complex tags and loose boundary division to obtain a wide range of causal categories and the complete causal semantics, but reduces the purity of causal semantics. However, the method of phrase boundary is s a compromise.

5 The experiment of optimal causal sequence labeling method

According to the Section 4, we conclude that most methods are balanced in tag feature complexity and causal scope breadth, so there is no perfect label method. We adopt the advantages and take out the disadvantages of the existing labeling methods, summarize several candidate labeling sequences, exploring the optimal label method through experiments and give the suggestions for method selection.

5.1 Experiment data

Due to the complexity of causal tags in the clause subscript method, it is difficult to train the feature, we adopt the tags in the phrase boundary method. Since "clause"unit is too extensive, losing the original meaning for causality extraction, the units of "core word" and "phrase" are selected.

There are controversies in phrase: (1) whether the attributive or adjective modifying the phrase needs to be labeled. (2) whether function word such as "the" before the phrases needs to be labeled. (3) there is a dispute over marking the former part, the latter part or the whole "of". According to the controversies

Label method	Multiple Causality	Embedded-Causality	Cross-Causality	Causal conjunction
Core Word	Location and quantity of tags C and E	Making rules to specify existing tag	×	×
Phrase Boundary	Location and quantity of tags B-C and B-E	Introducing new special tag "Emb"	×	×
Clause Subscript	Introducing subscript	Introducing compound label with subscript	Introducing subscript	Tag CC
Event sequence Blocks	Quantity of tags C and E in causality blocks	×	×	×

Table 4: Summary for solving special causality, "×" means the label method cannot solve the problem

Label method	Tag complexity	Tag semantic clarity	Causal type range	Label boundary strictness	Causal semantics completeness	Introducing no causal Semantics
CW	Easy	Fuzzier	Narrower	Strict	Incomplete	Slight
PB	Medium	Medium	Medium	Medium	Medium	Medium
CS	Difficult	Clear	Extensive	Loose	Complete	Heavy
ESB	Medium	Fuzzier	Narrow	Strict	Medium	Slight

Table 5: Summary for four labeling methods. The tag semantic clarity refers to whether the tag explicitly expresses the causality semantics such as embedded-causality, causal correspondence, etc. The "CW", "PB", "CS", "ESB" are the abbreviations of the core word, phrase boundary, clause subscript and event sequence block. The degree of comparison goes deeper to the prototype

in phrase, several candidate labeling sequences are proposed based on phrase boundary method. The details of are in Appendix A.

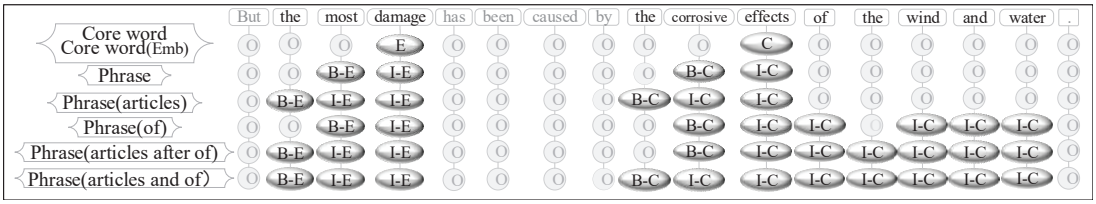

Figure 5: Example of a candidate causal labeling method

In addition, the SCIFI dataset described in Section 3.1 is used as the basic experimental data, 7 candidate labeling sequences are labeled respectively to form 7 sets of experimental data with the same text and different labels, forming the new dataset E-SCIFI(Extended SCIFI) for labeling sequence method.

5.2 Experiment content

We conduct labeling methods and comparative experiments which includes the 7 candidate labeling methods in Section 5.1. The basic model Bi-LSTM+CRF proposed in Huang et al. (2015) is selected to reduce the influence of the model itself on the experimental results. For the parameter setting, all the word vector dimension in our paper is 300. We train the model for 200 epochs with the learning rate of 0.5. And the optimizer is Adam.

Fine-grained evaluation criteria Take the sentence as unit, judge whether the causal extraction is correct according to the labeling sequence. If the labels of all the words in the sentence are correct, the causal relation extraction of the sentence is correct, including as follows: (1) the words and boundary of cause and effect extract correctly. (2) the causality is in the right direction. (3) cause and effect are extracted simultaneously. (4) In the case that multiple causality meet the above three conditions at the same time, the extraction is correct. The accuracy of sequence labeling is calculated as follows:

$$\text{accuracy} = \frac{m}{M} \times 100\% \tag{1}$$

where m is the number of correct sentences, and M is the total number of sentences.

Label Methods	C		E		Emb		O
	B-C	I-C	B-E	I-E	B-Emb	I-Emb	
Core word	1591		1527		0		22252
Core word(Emb)	1548		1510		60		22252
Phrase	1548	882	1510	753	60	39	20578
Phrase(articles)	1548	1647	1510	1461	60	71	19073
Phrase(of)	1548	1388	1510	1215	60	60	19389
Phrase (articles after of)	1548	1562	1510	1491	60	60	19139
Phrase (articles and of)	1548	2346	1510	2203	60	93	17610

Table 6: Label statistics of candidate causal labeling methods in E-SCIFI.

Coarse-grained evaluation criteria Take the label as unit, evaluate the value of F1 for each tag. The weighted sum of "B-X" and "I-X" tags in proportion to the number is used to measure the X-F1.

$$X - F1 = \frac{n_B}{n_B + n_I} \times f_B + \frac{n_I}{n_B + n_I} \times f_I \tag{2}$$

Where n_B, n_I are the number of tags "B-X" and "I-X", f_B, f_I are the corresponding F1 value.

5.3 Experimental results and analysis

Label Methods	Overall Accuracy(%)	C-F1(%)		E-F1(%)		Emb-F1(%)		O-F1(%)
		B-C	I-C	B-E	I-E	B-Emb	I-Emb	
Core word	62.30	83.96		88.46		-		98.16
Core word (Emb)	60.21	83.69		86.38		33.33		98.05
Phrase	57.89	79.90		80.08		11.55		96.77
		85.27	70.48	83.05	74.11	19.05	0.00	
Phrase(articles)	60.21	82.16		84.73		0.00		96.22
		84.85	79.63	84.76	84.70	0.00	0.00	
Phrase(of)	48.42	79.90		80.91		35.90		95.57
		87.25	71.71	84.76	74.29	33.33	38.46	
Phrase(articles after of)	50.53	77.08		76.89		19.29		94.72
		86.14	68.11	82.35	71.37	14.29	10.00	
Phrase(articles and of)	52.36	81.66		79.07		17.63		94.14
		85.65	79.02	80.92	77.81	28.57	12.50	

Table 7: Experiment results for the labeling methods. The bold indicates the optimal result

Causal unit analysis The accuracy for core word is higher than phrase, the number and type of label are in a small amount in core word method, the features are simple and easy to learn.

"Emb" tag analysis: The number of tag "Emb" in the data is too small to train accurate features, which reduces the accuracy of whole experiment, reducing 2.09% compared with the core word method without "Emb". It can be seen that the embedded-causality words are clearly marked semantically by introducing new tags, but the experimental effect is reduced. Therefore, "making rules to specify existing label" can be used as a compromise to sacrifice semantic clarity of tags for training feature accuracy.

Causal tag analysis As the largest number of entities labelled "O", the feature is obvious and the F1 value is the highest. In the most label methods, the F1 value of tag "E" is higher than that of tag "C", the ability of the model to identify the effect entity is higher than that of the cause entity.

Boundary tag analysis In all phrase labeling methods ("Emb" tag of phrase(of) except), the F1 value of "B-X" tag is higher than that of "I-X", so the ability to identify the beginning boundary of causal phrase is higher than that of the end(middle). Most of the phrase boundary disputes are over the setting of the phrase end boundary, while the phrase start boundary disputes are less.

Phrase labeling analysis (1) with articles is better than without articles: the article feature (the, an, a corresponding word vector) can explicitly mark the starting position of causal phrase even if it does not have causal semantics. (2) label part of "of" is better than the whole "of": labeling the whole "of" has less controversy in which part of "of", however, the text of causal phrase is too long, introducing boundary division disputes and adding more useless feature. (3) continuous labeling sequence is better

than discontinuous sequence: The "phrase(of)" method has discontinuity in the labeling sequence, which destroys the continuity of causal entity and increases the difficulty of extracting features.

With the combine of the accuracy of labeling sequence and the F1 value of the tag, and only the experimental effect was used as the criterion, the optimal labeling method was ranked as: core word, core word(Emb), phrase(article), phrase, phrase(articles and of), phrase(articles after of), phrase(of).

5.4 Analysis of optimal causal sequence labeling method

If the research has no special requirements, the labeling rank in Section 5.3 can be used as the priority of the selection of causal labeling method. However, in fact, the selection of labeling method still needs to be combined with the research purpose and experimental data.

Under the situation that there is no causal semantic integrity requirement, and the embedded-causality extraction is not the focus task, the "core word" labeling method is preferred. However, if the experiment focuses on the study of embedded-causality, "core word (Emb)" should be selected.

If the research has certain requirements on causal semantic integrity, the optimal labeling method should be selected from the phrase unit. If there is no requirement for causal purity, the "phrase(article)" labeling method is preferred, otherwise, phrase method can be adopted by sacrificing a little accuracy.

6 Conclusions

Causal relationship extraction is still a new research field with no publicly evaluated dataset and fixed labeling method, which are the basis of all research and one of the important factors to hinder the progress of causality extraction. We summarize and analyzes the defects in the construction of the experimental data and the disputes over the labeling method for causality in all aspects, so as to serve as a reference for the further research on causality. We also explore the optimal labeling method of causal sequence through experiments, and puts forward suggestions for the selection of labeling methods. In addition, we explore the existing research on causality from multiple perspectives, summarizing the causality extraction related concepts (See the Appendix B for the detailed table).

Acknowledgements

This work is sponsored by the National Natural Science Foundation of China (61976103, 61872161), the Scientific and Technological Development Program of Jilin Province (20190302029GX, 20180101330JC, 20180101328JC). All opinions, findings, conclusions and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Eduardo Blanco, Nuria Castell, Dan Moldovan. 2008. Causal Relation Extraction. *In Pro-ceedings of the 6th International Conference on Language Resources and Evaluation*, 310-313, Marrakech,Morocco.

Danushka Bollegala, Simon Maskell, Richard Sloane, et al. 2018. Causality patterns for detecting adverse drug reactions from social media: text mining approach. *JMIR public health and surveillance*, 4(2): e51.

Arantza Casillas, Alicia Pérez, Maite Oronoz, et al. 2016. Learning to extract adverse drug reaction events from electronic health records in Spanish. *Expert Systems with Applications*, 61: 235-245.

Tirthankar Dasgupta, Rupsa Saha, Lipika Dey, et al. 2018. Automatic Extraction of Causal Relations from Text using Linguistically Informed Deep Neural Networks. *In Proc of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. 306-316, Melbourne, Australia.

Quang Do, Yee Seng Chan, Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 294-303, Edinburgh, UK.

Chong Feng , Liqi Kang , Ge Se, et al. 2018. Causality Extraction With GAN. *Acta Automatica Sinica*, 44(5): 811-818.

Juliane Fluck, Sumit Madan, Sam Ansari, et al. 2016. Training and evaluation corpora for the extraction of causal relationships encoded in biological expression language (BEL). *Database*, Volume 2016, baw113.

Jianfeng Fu , Zongtian Liu , Wei Li , et al. 2011. Event Causal Relation Extraction Based on Cascaded Conditional Random Fields. *Pattern Recognition and Artificial Intelligence*, 24(4):567-573.

Daniela Garcia.1997. COATIS, an NLP system to locate expressions of actions connected by causality links. In *Proc of the 10th European Workshop on Knowledge Acquisition, Modeling and Management*, 347-352, Catalonia, Spain.

Roxana Girju. 2003. Automatic Detection of Causal Relations for Question Answering. In *Proceedings of the 41st ACL Workshop on Multilingual Summarization and Question Answering*, 2003: 76-83, Sapporo, Japan.

C.W.J. Granger. 1988. Some recent development in a concept of causality. *Journal of econometrics*, 39(1-2): 199-211.

Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, et al. 2012. Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics*, 619-630, Jeju Island, Korea.

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, et al. 2014. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*(Volume 1: Long Papers), 987-997, Kyoto, Japan.

Hashimoto C, Kentaro Torisawa, Julien Kloetzer, et al. 2015. Generating event causality hypotheses through semantic relations. *Twenty-Ninth AAAI Conference on Artificial Intelligence*. Austin, Texas, USA.

Christopher Hidey, Kathy McKeown. 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), 1424-1433, Berlin, Germany.

Zhiheng Huang, Wei Xu, Kai Yu.2015. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508. 01991.

Ashwin Ittoo and Gosse Bouma. 2011. Extracting explicit and implicit causal relations from sparse, domain-specific texts. In *International Conference on Application of Natural Language to Information Systems*, 52-63, Berlin, Heidelberg.

Dongyeop Kang, Varun Gangal, Ang Lu, et al. 2017. Detecting and explaining causes from text for a time series event. arXiv preprint arXiv:1707.08852.

Humayun, Kayesh, Md Saiful Islam, Junhu Wang. 2019. On Event Causality Detection in Tweets. arXiv preprint arXiv:1901.03526, 2019.

Christopher S. G. Khoo, Jaklin Kornfilt, Oddy Robert N, et al. 2003. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing, *Literary and Linguistic Computing*, 13(4): 177-186.

Christopher S. G. Khoo, Syin Chan, Yun Niu. 2000. Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 336-343, Hong Kong.

Murat Kocaoglu, Christopher Snyder, Alexandros G. Dimakis, et al. 2017. Causalgan: Learning causal implicit generative models with adversarial training. arXiv preprint arXiv:1709.02023.

Canasai Kruengkrai, Kentaro Torisawa, Chikara Hashimoto, et al. 2017. *AAAI Conference on Artificial Intelligence, 3466-3473, New York.*

Pengfei Li, Kezhi Mao. 2019. Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. Expert Systems with Applications, 115: 512-523.

Zhaoning Li, Qi Li, Xiaotian Zou, et al. 2019. Causality Extraction based on Self-Attentive BiLSTM-CRF with Transferred Embeddings, arXiv preprint arXiv:1904.07629, 2019.

Zhiyi Luo, Yuchen Sha, Kenny Q. Zhu, et al. 2016. Commonsense causal reasoning between short texts. In *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Paramita Mirza. 2014. Extracting temporal and causal relations between events. In *Proceedings of the ACL 2014 Student Research Workshop*, 10-17, Baltimore, Maryland USA.

Paramita Mirza, Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*: Technical Papers, 2097-2106, Dublin, Ireland.

G Nordon, G Koren, Varda Shalev, et al. 2019. Building Causal Graphs from Medical Literature and Electronic Medical Records. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33: 1102-1109, New York.

Osonde Osoba, Bart Kosko. 2019. Beyond DAGs: modeling causal feedback with fuzzy cognitive maps. arXiv preprint arXiv:1906.11247.

Michael J. Paul. 2017. Feature selection as causal inference: Experiments with text classification. In *Proceedings of the 21st Conference on Computational Natural Language Learning*,163-172, Vancouver, Canada.

Jiangnan Qiu, Liwei Xu, Jie Zhai, et al. 2017. Extracting Causal Relations from Emergency Cases Based on Conditional Random Fields. *Procedia Computer Science*, 112: 1623-1632

Bilal Qureshi, Faisal Kamiran, Asim Karim, et al. 2016. Causal discrimination discovery through propensity score analysis. arXiv preprint arXiv:1608.03735.

Kira Radinsky, Sagie Davidovich, Shaul Markovitch. 2012. Learning causality for news events prediction. In *Proceedings of the 21st international conference on World Wide Web*, 909-918, New York.

[Raja et al.2013] Kalpana Raja, Suresh Subramani, Jeyakumar Natarajan. 2013. PPInterFindera mining tool for extracting causal relations on human proteins from literature. Database, Volume 2013, bas052.

Fabio Rinaldi, Tilia Renate Ellendorff, Sumit Madan, et al. 2016. BioCreative V track 4: a shared task for the extraction of causal network information using the Biological Expression Language. Database, Volume 2016, baw067.

Jawad Sadek, Farid Meziane. 2016. Extracting Arabic causal relations using linguistic patterns. ACM *Transactions on Asian and Low-Resource Language Information Processing* (TALLIP), 15(3): 1-20.

Tharini N. de Silva, Xiao Zhibo, Zhao Rui, et al. 2017. Causal Relation Identification Using Convolutional Neural Networks and Knowledge Based Features. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 11(6): 697-702.

Yizhou Sun, Kunqing Xie, Ning Liu, et al. 2007. Causal relation of queries from temporal logs. In *Proceedings of the 16th international conference on World Wide Web*. 2007: 1141-1142, New York.

Bögel, Tina, Hautli-Janisz, Annette; Sulger, et al. Automatic detection of causal relations in German multilogs. 2014. *14th Conference of the European chapter of the association for computational linguistics*. 2014: 20-27, Stroudsburg.

White, Peter A. Ideas about causation in philosophy and psychology. Psychological bulletin, 108(1): 3.

Jinyoung Yeo, Gengyu Wang, Hyunsouk Cho, et al. 2018. Machine-Translated Knowledge Transfer for Commonsense Causal Reasoning. *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

Sendong Zhao, Ting Liu, Sicheng Zhao, et al. 2016. Event causality extraction based on connectives analysis. *Neurocomputing*, 173(P3):1943-1950.

Sendong Zhao, Quan Wang, Sean Massung, et al. 2017. Constructing and embedding abstract event causality networks from text snippets. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 335-344, New York.

A Candidate labeling sequences

In our experiments, the details of proposed candidate labeling sequences are as follows:

- **Core word**: The causal labeling method of core word in Section 4.1.
- **Core word**: The core word methods with new tag "Emb" to label embedded-causality.
- **Phrase**: Label the attributive adjective and part of "of", ignore the articles.
- **Phrase(articles)**: Label the attributive adjective, articles and part of the "of".
- **Phrase(of)**: Ignore all articles, label the whole "of", with a causal break in the labeling sequence.
- **Phrase(articles after of)**: Ignore the articles before "of" and label the whole "of".
- **Phrase(articles and of)**: Label all articles and the whole "of".

B Causality extraction related concepts

Concept	Summary
2 special multiple causality	Embedded-Causality, Cross-Causality
3 research methods	Text classification, Relation extraction, Sequence labelling
3 steps of causal sequence label	Setting the causal semantic labels, Establishing the causal labeling units, Special causality processing
4 causality units	Core word, Phrase, Clause, Event
4 label methods	Core word, Phrase boundary, Clause subscript, Event sequence block
6 public dataset	SemEval, Causal-TimeBank, altlex, CEC, SCIFI, ESC

Table 8: Summary of the causal relationship extraction related concepts.