Knowledge and Information Systems
https://doi.org/10.1007/s10115-022-01781-7

REGULAR PAPER

A storytree-based model for inter-document causal relation extraction from news articles

Chong Zhang[1] · Jiagao Lyu[1] · Ke Xu[1]

Received: 3 March 2021 / Revised: 9 October 2022 / Accepted: 16 October 2022

Abstract
With more and more news articles appearing on the Internet, discovering causal relations between news articles is very important for people to understand the development of news. Extracting the causal relations between news articles is an inter-document relation extraction task. Existing works on relation extraction cannot solve it well because of the following two reasons: (1) most relation extraction models are intra-document models, which focus on relation extraction between entities. However, news articles are many times longer and more complex than entities, which makes the inter-document relation extraction task harder than intra-document. (2) Existing inter-document relation extraction models rely on similarity information between news articles, which could limit the performance of extraction methods. In this paper, we propose an inter-document model based on storytree information to extract causal relations between news articles. We adopt storytree information to integer linear programming (ILP) and design the storytree constraints for the ILP objective function. Experimental results show that all the constraints are effective and the proposed method outperforms widely used machine learning models and a state-of-the-art deep learning model, with F1 improved by more than 5% on three different datasets. Further analysis shows that five constraints in our model improve the results to varying degrees and the effects on the three datasets are different. The experiment about link features also suggests the positive influence of link information.

Keywords Relation classification · News article · Causal relation · Constraint

1 Introduction

News keeps people informed about events happening around the world. With the increase in the amount of information, the amount of news on news websites has exploded. Understanding the relation between various news articles allows us to better sort out the development of events and has a deeper understanding of various news. Therefore, it is meaningful and

✉ Ke Xu
kexu@buaa.edu.cn

[1] State Key Lab of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing, China

Published online: 03 November 2022

necessary to extract the relation between news articles. In recent years, there has been more research on news analysis [18, 20, 37, 42].

Relation extraction (RE) is an important natural language processing task of identifying the relation between entities. The extracted relations can be used as entity links to build entity graphs or networks, which are fundamental for many downstream tasks, such as entity graph analysis [47], graph-based question-answering [6], and decision support system [24]. In recent years, more and more attention has been devoted to the extraction of the causal relation, and many methods are proposed for this task. Extracting causal relation helps to identify cause-effect entity pairs which provide essential information for natural language understanding.

Unlike the intra-document task identifying the relations of trivial events, the inter-document RE task focuses on the relation of documents, the documents can be academic papers, news articles, etc. For inter-document RE tasks, the intra-document RE model is difficult to work is because for the following reasons: (1) Documents are much more complicated than entities. Taking news articles as an example, a 500-word news article is dozens of times longer than an intra-document task data. If we use the intra-document task model based on word embedding, the complexity of the model will become higher and difficult to converge, which makes the model difficult to train. (2) In the intra-document task, there will be connectives between entities or sentences (such as 'because' and 'so'). The connectives can help the model determine the relation between entities. But in the inter-document task, there is no direct connection between articles. Therefore, the intra-document task models with connectives as an important feature are difficult to be effective. In this study, we focus on the causal relation between news articles. The usual inter-document RE methods use classifiers to determine the relation based on the similarity between two documents [13, 21], so the performance of these methods heavily relies on the accuracy of similarity. In two news articles with causal relation, their characters, entities, and other elements are similar. However, it is also possible that two similar news articles are not causally related. For example, news articles D_a and D_b describe two cases committed by one murderer in the same place, which are not causally related. News article D_c describes the arrest of this murderer, apparently causally related to both D_a and D_b. But the similarity between D_a and D_b is higher than the similarity between D_a and D_c. In real data, there are many similar examples, so the classifier is likely to misclassify.

Some previous relation extraction models [32, 45] organize events into a storyline. The storyline is useful for temporal relations because of its linear structure, but it is not suitable to demonstrate complex causal relations between events. In recent studies, the storytree [19] structure is successfully employed to extract events from news text. The storytree is a tree structure to visualize the news corpora, and each document belongs to a story that is a node in the storytree. When news documents are organized and clustered into several stories, the storytrees are used to show the evolution of the stories. Intuitively, the tree structure could provide features that benefit our causal relation extraction. Therefore, we use the relative position of nodes to identify the relation between two events in this paper. The storytree not only clusters and organizes news events into stories, but also provides extra features of story links and positions for causal relation extraction, which is why the storytree is effective in our research.

In this study, we propose a causal relation extraction model. First, we crawled news reports separately from the three news websites. Each news article is saved as a file. The relevant information of the news, including the release time and content, is stored. We use data augmentation technology to enlarge the datasets and label the data. We end up with three datasets containing over 1000 news articles. We employ EventX [19] to cluster the

news articles and generate the storytree of each dataset and use Integer Linear Programming (ILP) to model the causal relation between two news articles by designing constraints and adjusting the objective function. We also take storytree information into the ILP framework to formulate constraint learning. This not only improves the base classifier of our model by more than 20%, but our model also outperforms the widely used classification models by more than 5%.

The contributions of our model can be summarized as follows:

1. We propose a novel inter-document causal relation extraction model for news articles. Our task is different from the intra-document RE task because news articles are longer and more complex than the entities in the intra-document RE task. So extracting the causal relation between news articles is a new and meaningful task. Our model uses advanced data processing methods and models and achieves good performance.
2. We are the first to use the storytree information in the causal relation extraction model and achieve good performance. We use the storytree information to design five constraints in the Integer Linear Programming model and prove the effectiveness of each constraint through an ablation experiment.
   The rest of this article is organized as follows. First, Sect. 2 gives an overview of the related work about relation extraction and timeline generation. Next, Sect. 3 introduces the formal definition of this task. Section 4 describes the structure and detail of our model. The datasets and experimental results are presented in Sect. 5. Section 6 analyzes and discusses the results further. Finally, we end up with conclusions in Sect. 7.

2 Related work

Intra-document relation extraction is to extract various relations between entities in a document. It can be divided into sentence-level RE, document-level RE, etc. In the last decade, many intra-document relation extraction approaches have been proposed. Some methods are about sentence-level RE, which is extracting relational facts from a single sentence. With the development of neural models, various models have been proposed to encode relational patterns of sentence-level entities. Despite the state-of-the-art performance [8, 16, 31], sentence-level RE ignores the entity relations in multiple sentences. Therefore, in recent years, more studies focus on document-level RE [41]. The new dataset constructed from Wikipedia and Wikidata called DocRed [46] lays a good foundation for document-level RE. BERT-based models are proposed and achieve good performance on DocRED [14, 26]. Hierarchical Inference Network (HIN) is proposed to make full use of the abundant information and achieves state-of-the-art performance on DocRED [40].

Inter-document RE is the study of more macroscopic events, and events are represented by the whole document. Previous work has dealt with inter-document relations in two ways. The first way is to use feature-based models. A semi-supervised classification algorithm learns from dissimilarity and similarity information on labeled and unlabeled data; this approach uses a novel graph-based encoding of dissimilarity and can handle both binary and multi-class relation classification [13]. Social relations and text similarities are incorporated into building microblog-microblog relations [21]. MuReX is a complete system to discern salient connections and facts from a set of related documents [37]; ILP is used to ensure the relevance of user queries to facts, but this study focuses on information extraction and visualization. HINT [5] introduces the shared 'anchor texts' to connect the comparative news, and then, two similarity matrices, as well as a transition matrix for cross-text-source knowledge transfer,

are constructed for clustering the news to get 'comparative' information between news. Some researches focus on Improving the Quality of features. A feature-based model of two parts is proposed, and the first part is an extension of a hierarchical topic model that induces background, relation specific, and argument-pair-specific feature distributions. The second part is a perceptron trained to match an objective function that enforces constraints semantics and noises [35]. The second way is to use deep learning methods. Shallow and deep learning are both used for event-relatedness classification [15]; they compare the performance of shallow learning methods and deep learning approaches based on long short-term memory (LSTM) recurrent neural network. Other studies are mainly about the implicit relation of discourse (IDRR), which is shorter than news data. The causal relation is one of the implicit relations between discourses. A stacking neural network model is proposed in which a convolutional neural network (CNN) is utilized for sentence modeling and a collaborative gated neural network (CGNN) is for feature transformation [30]. A hierarchical multi-task neural network with a conditional random field layer (HierMTN-CRF) is proposed for multi-level IDRR [44] and achieves significantly better and more consistent results over several competitive baselines on multi-level IDRR. In our study, the news document is much longer than the discourse, which greatly increases the training cost of the deep learning model, so we use a feature-based method to solve the causal relation extraction of news documents.

In recent years, constrained conditional models (CCMs) [7] are proposed to augment the learning of conditional models with declarative constraints. Integer linear programming (ILP) is used as the inference framework of CCM in RE tasks [12, 27]. In this paper, ILP is also used to optimize the classification model by designing constraints related to the storyline information in the data.

Storyline extraction is aimed at summarizing the development of certain related events. Many studies of storyline extraction are from text. Yang et al. [45] combine the similarity information of events, the distance of events, temporal sequence, and the distribution of documents along the timeline to score the relation of two events; they use a directed acyclic graph (DAG) to model the evolution of events. Ls et al. [22] construct an Event-Oriented Similarity Graph to represent the relationship among retrieved event news documents and develop a community detection algorithm to segment sub-events which are consequently chained into a cohesive event line. Radinsky and Horvitz [32] clusters text and uses entity entropy to generate storylines. Besides, an approach is based on discrete dynamic topic modeling and Hidden Markov Model for event detection and tracking; then, the events that would persist in the next time slice are predicted [23]. To improve the clustering accuracy, more multi-stage clustering models are proposed. Shahaf et al. [36] propose a two-stage clustering process and use word clustering to retain most of the mutual information between words and documents. Another two-stage approach consisting of date selection and date summarization is proposed to build timelines [39]. Story Forest [19] is proposed to cluster documents into events. It connected related events in growing trees to tell evolving stories. The two-layer clustering approach called EventX is based on both keyword graphs and document graphs. Story Forest accurately identifies events and organizes news text into a logical structure. In this paper, we used EventX to cluster the data and build the storytree of the data.

3 Task definition

In this study, the objective is to extract the correct causal relation news article pair from the datasets. We model inter-document causal relation extraction as a binary classification task of the triplet. Therefore, the inter-document causal relation extraction task is defined as follows: given a set of news articles $D = \{d_1, d_2, \ldots, d_n\}$. The model aims to correctly output a set of triples Y that indicate whether there is a causal relation between two news articles:

$$Y = \{(d_i, d_j, r) | d_i, d_j \in D, r \in \{0, 1\}\} \tag{1}$$

where $r = 1$ means that there is a causal relation between news article d_i and d_j. Note that $(d_i, d_j, r) = (d_j, d_i, r)$, because we consider that causal relation between two news article is undirected, the news article occurs earlier is the source and the other is the target.

In this task, each news article is a full news report, and each article is associated with a template that consists of basic information about the news, such as ID, type, time, etc. Some information is mandatory, e.g. ID, type, content, time, and others are optional, e.g., the location of the link contained in the news report. According to Wikipedia's definition of the causal relation[1], causality is influenced by which one event, process, state, or object (a cause) contributes to the production of another event, process, state, or object (an effect) where the cause is partly responsible for the effect, and the effect is partly dependent on the cause. Taking news articles as an example, news articles a have a causal relationship with news articles b, which means that the generation of a leads to the generation of b (or vice versa), one happens before and the other happens in Behind. What happens in front is "cause", what happens behind is "effect". If the "cause" does not happen, then the "effect" does not happen either. Figure 1 provides some examples of causality in news articles. Article and news article have the same meaning in the following text.

At present, there is little research on the task of inter-document causal relation extraction. We have given a clear task definition and proposed a new model based on the storytree for this task.

4 Model

The structure of the model is shown in Fig. 2. We first preprocess the news articles by removing the stop words and punctuations and building the vocabulary list of each article. Then, we extract the keywords of each article. All the keywords are fed into the storytree generation part; we cluster the keywords through the EventX algorithm and then build the storytree. In addition, we extract the features of each article and use Logistic Regression (LR) as the model classifier. In the ILP model, storytree information is used to design constraints. We design 2 basic constraints and 5 types of constraints based on storytrees.

4.1 Preprocessing

For each news article, the length of the text is very long and the content is complex, so preprocessing is very important. First, we clean the text to remove punctuation, symbols, etc., in the text. Then, we perform word segmentation, entity extraction and other operations on the text through the spaCy toolkit. SpaCy is a library for advanced Natural Language Processing in Python; it features state-of-the-art speed and neural network models for tagging, parsing,

[1] https://en.m.wikipedia.org/wiki/Causality.

document A	
ID	*3*
TIME	*'January 27, 2020'*
TITLE	*'Kobe Bryant Killed in Helicopter Crash at 41'*
TEXT	*'Kobe Bryant was killed in a helicopter crash in Calabasas Sunday morning, a source confirms to PEOPLE. The NBA legend, 41, was reportedly traveling with at least three other people in his private helicopter when it went down, according to ...'*
...	

document B	
ID	*8*
TIME	*'February 28, 2020'*
TITLE	*'Kemba Walker explains keeping No. 8 to honor Kobe Bryant: 'We want to keep his legacy going'*
TEXT	*'Whenever Kemba Walker looks down at the No. 8 on his jersey, he'll be reminded of the Mamba Mentality. While others in the NBA switched their jersey numbers in the wake of Kobe Bryant's tragic death, Walker instead decided to...'*
...	

document C	
ID	*24*
TIME	*'March 12, 2020'*
TITLE	*'State Department warns Americans to reconsider traveling abroad due to coronavirus'*
TEXT	*'The U.S. State Department warned Americans late Wednesday to avoid traveling abroad, in response to a coronavirus outbreak that's reached pandemic status.Also Wednesday, the...'*
...	

Fig. 1 Examples of causal relations. The occurrence of document *A* (the death of Kobe) led to the occurrence of document *B* (the tribute to Kobe), *A* is the "cause," and *B* is the "effect." Both *A* and *B* have no causal relation with document *C* because *C* does not describe the same story with *A* and *B*

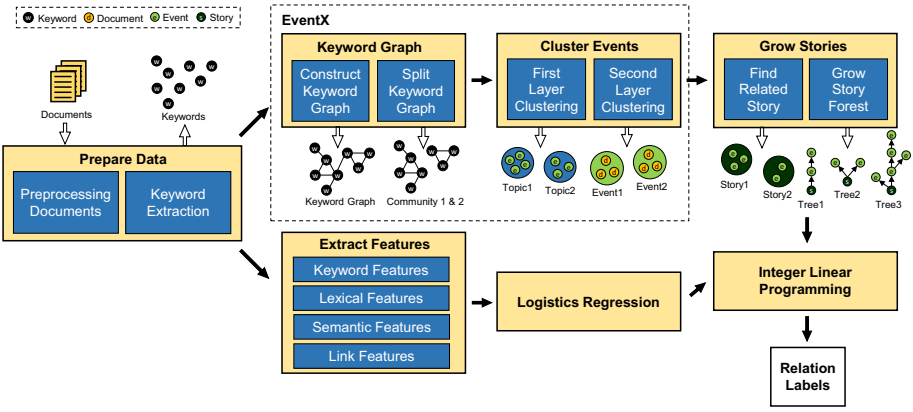


Fig. 2 An overview of structure of the model

named entity recognition, and text classification. We obtain all entities in the news article d through entity extraction and count the occurrence frequency of the entities. Then, we filter out the entity set $\text{En} = \{e_1, e_2, ..., e_n\}$ whose entity occurrence frequency is greater than 1, where e_i is the ith entity in the text. Additionally, we obtain the vocabulary list V of document d and compute the TF-IDF value for each word in the vocabulary. TF-IDF [38] is used to assess the importance of a word to a document set or a document in a corpus. TF-IDF is the product of two statistics, term frequency, and inverse document frequency. The TF-IDF value for word v is calculated by:

$$\text{tfidf}_v = \frac{f_v}{\sum_{v_i \in V} f_{v_i}} * \log \frac{|D|}{|\{d \in D : v \in d\}|}, \tag{2}$$

where V is the vocabulary list of document d, D represents the set of documents, and f_v refers to the frequency of v in document d. The first part of the equation is the term frequency (TF), and the second part is the inverse document frequency (IDF). We take the top m words with the highest TFIDF value in each article as $\text{Kw} = \{v_1, v_2, ..., v_m\}$. Finally, we merge the entity set En with the Kw as the keyword set K_d for document d, where $K_d = \text{En} \cup \text{Kw}$.

4.2 Storytree generation

Storytree generation consists of two parts. The first part is to perform 2-layer clustering with the EventX algorithm. The second part is building the storytree.

EventX EventX is a 2-layer clustering algorithm based on keyword graphs and document graphs [19]. At the beginning of the algorithm, a keyword co-occurrence graph $G = (V, E)$ is first constructed for the news article corpus D. Each node in G is a keyword extracted by Sect. 4.1. The edges is $E = \{e_{ij} \mid f(k_i, k_j, D) > \delta_e, \Pr(k_i \mid k_j) > \delta_p, \Pr(k_j \mid k_i) > \delta_p\}$, where k_i is the keyword of a news article in corpus D, $f(k_i, k_j, D)$ indicates the times of co-occurrence of k_i and k_j in one news article, $\Pr(k_i \mid k_j)$ is calculated by Eq. 3, which refers to the conditional probabilities of the occurrence of k_i and k_j, $\text{DF}_{i,j}$ represents the number of news articles that contain both keyword k_i and k_j, and DF_j is j is the document frequency of keyword k_j. δ_e (we set $\delta_e = 2$ in out experiments) and δ_p (we use 0.05) are the thresholds.

$$\Pr(k_i \mid k_j) = \frac{\text{DF}_{i,j}}{\text{DF}_j}. \tag{3}$$

In the first layer of clustering, keywords in all news articles are used to form a keyword graph and segment the graph. For each keyword subgraph, a subset of the corpus that is highly relevant is retrieved. In the second layer, news articles under each keyword subgraph form an article graph, and the connected edges represent two articles telling about the same event. Community detection is performed on the article graph. Each news article subgraph community represents an event. The details of the algorithm are shown in Algorithm 1. EventX gathers relevant news articles into one category through the two-layer clustering, and each news article community can be understood as a story. Based on these stories, we can construct storytrees for them.

Growing stories After event clustering is obtained, the different event nodes are inserted into the existing storytree or form new stories to form the storytree. We identify the related storytree of each event node by calculating the average similarities of news articles in the

Algorithm 1 EventX Clustering Algorithm

Input: A set of news articles $D = \{d_1, d_2, ..., d_{|D|}\}$, with extracted keywords described in Sect. 4.1.
Output: A set of events $E = \{E_1, E_2, ..., E_{|E|}\}$.
1: Construct a keyword co-occurrence graph G of D based on all news articles' keywords.
2: Split G into a set of strongly connected keyword communities $C = \{C_1, C_2, ..., C_{|C|}\}$ by a community detection algorithm[29].
3: **for** each keyword community $C_i, i = 1, ..., |C|$ **do**
4: Calculate the cosine similarity between the TF-IDF vector of each news article and the keyword community, compare it to a threshold δ. The news articles subset D_i that is highly relevant to the keyword community is retrieved.
5: Connect news article pairs in D_i with the similarity between news articles to form a document relationship graph G_i^d.
6: Based on the community detection algorithm, G_i^d, is split into a set of news article communities $C_i^d = \{C_{i,1}^d, C_{i,2}^d, ..., C_{i,|C|}^d\}$.
7: **end for**

event node and existing storytrees. The different event nodes are inserted into the existing storytree or form new stories to form the storytree.

The storytrees generated by two-layer clustering sort out the logical relation of news articles and help us understand the position of each article in the story, so the storytree information is significant for our model. Storytree information is employed to design constraints in the ILP model.

4.3 Similarity measurement

The similarity between news articles is an important feature; cosine similarity, best match 25 (BM25) [17, 34] and Jaccard similarity are widely used in existing bag-of-words (BoW) inter-document similarity measures. In this paper, we use a new measure for inter-document similarity measurement called Sp [1]. The similarity of two news articles a and b can be formulated by Eq. 4:

$$s_{\mathrm{sp}}(a,b) = \frac{1}{|T_a \cup T_b|} \sum_{t_i \in T_a \cap T_b} \log \frac{N}{|\{c \in D : \min(a_i, b_i) \le c_i \le \max(a_i, b_i)\}|} \tag{4}$$

where T_a, T_b is the set of keywords in news article a and b, D is the collection of news articles, N is the number of news articles in D, a_i, b_i and c_i represents the occurrence frequency of keyword k_i in news article a, b and c, respectively. This measure does not need term weighting, and we can also know from the formula that it is related to traditional Jaccard similarity and IDF term weighting. In [1], they identify the shortcomings of the underlying assumptions of term weighting in the inter-document similarity measurement task. In their paper, some examples were shown to prove that only judging the importance of the weight of terms in the BoW vector of one article can be counter-productive in inter-document similarity measurements. In inter-document similarity measurement experiments, Sp has the best performance over BM25, cosine similarity, and Jaccard similarity, so we use Sp in the feature extraction part and the second layer clustering of events clustering part to measure the similarity of news articles in this paper.

4.4 Feature extraction

We have extracted the keywords from the news articles before we build the feature set. In our model, the keyword set is like a summary of the news described in the article, so the keyword set should contain named entities and verbs. We have used spaCy to extract the named entities in the article and put the entities mentioned more than once into the keyword set. The verbs with high TF-IDF values in the article have also been put into the keyword set. The keyword sets are used to calculate the similarity between news articles. We use the same set of features for training all the classifiers.

Keyword features According to Sect. 4.1, the keywords of each news article have been extracted. A simple way to get the keyword level similarity information is to get the number of common keywords between news articles. We obtain the keyword features by calculating the number of words in the intersection of the two article keyword sets.

Lexical features We first use Sp to calculate the similarity of each pair of news articles. The headlines and first sentences of news articles tend to generalize the news and are used as features in document-level classifiers [19]. To capture more information about the similarity between news articles, we calculate the TF-IDF similarity of the title and the TF-IDF similarity of the first sentence.

Semantic features We train a Latent Dirichlet Allocation (LDA) [3] model based on the content of the article to get the LDA feature vector. Then, we calculate the LDA cosine similarity of two news articles as the semantic feature of our model.

Link features On news sites, we often see text with hyperlinks that redirect to other news articles. These links represent relations between news articles directly and accurately, so we consider them to be a feature of classifiers. Locations of texts with links are saved when we collected the news. We define lw_a is the word set of link text in article a, w_b is the keyword set of article b, a occurs later than b. The link characteristics of two articles are calculated by $|lw_a \cap w_b|$.

4.5 Logistic regression

The logistic regression (LR) model is the base classifier of our model using all the features. Compared with other classifiers, logistic regression is more suitable for binary classification, and the training speed is fast. In addition, the output of the model is a probability value, which is very convenient for the access of the following integer linear programming model. In LR model, the parameters of each feature are obtained by the gradient ascent method. For the binary classification problem coded by $y \in \{0, 1\}$, we use the sigmoid function to calculate the probability that each example is a positive example: $P(y = 1|x) = \exp(\beta_0 + \beta^T x_i)(1 + \exp(\beta_0 + \beta^T x_i))$, where $\beta^{'}$s are the parameters and x represents the features. The output of the LR model is input to the ILP system.

4.6 Storytree based causal RE using ILP

Integer linear programming (ILP) is a type of linear programming in which all decision variables are integers. In our research, the decision variable $x \in \{0, 1\}$ is because our task is a binary classification problem. The ILP system performs global inference through constraints

for resolving the causal relation between news articles. We first design two basic constraints based on news articles and then design five types of constraints based on the information of nodes and links. We define p_{ij} as the confidence score of the causal relation between article i and j obtained by the classifier, D is the set of news articles, and Dp represents the set of news article pairs. $x_{ij} \in \{0, 1\}$ represents the decision variable of the news article pair i and j in the ILP model. After the integer linear programming problem is solved, x_{ij} is the final classification result of the news article pair i and j.

Basic constraints One of the two basic constraints is that the system discourages causal relations between two news articles occurring at the same time. The other is that the system discourages causal relations between two news articles with very low similarity. The objective function is formulated with Eq. 5, and two basic constraints are formulated with Eqs. 6 and 7,

$$Y_{\text{Obj}} = \max \sum_{i \in D} \sum_{j \in D} [x_{ij} p_{ij} + \neg x_{ij}(1 - p_{ij})] \tag{5}$$

$$\forall (i, j) \in Dp, x_{ij} \leq t_{ij} \tag{6}$$

$$\forall (i, j) \in Dp, x_{ij} \leq \frac{s_{ij}}{k_b} \tag{7}$$

where t_{ij} is the absolute value of the time difference between news articles i and j, s_{ij} is the similarity of article i and j, $0 \leq k_b < 1$ is a hyperparameter to reduce the model's dependence on similarity.

Same node constraints (SMN) Each node of the storytree represents an event, and an event may include more than one news article. If two news articles are in the same node, they are probably about the same topic, and more likely to have a causal relation, so the system encourages two articles on the same node to have a causal relation. The constraint can be represented by Eq. 8:

$$\forall (i, j) \in Dp, x_{ij} \geq (1_{i \in N} \wedge 1_{j \in N})(k_{sn} - \frac{t_{ij}}{t_m}) \tag{8}$$

In the above equation, N is a node of the storytree, t_m is the maximum of the absolute value of the time difference between news articles in the dataset, and k_{sn} is the hyperparameter. We define $1_{\text{condition}} = 1$ if condition is true, otherwise $1_{\text{condition}} = 0$. In Eq. 8, $1_{i \in N} \wedge 1_{j \in N}$ means that this constraint is valid only if news articles i and j are in the same node; otherwise, the equation is always established and the constraint is invalid.

Root node constraints (RTN) The root node of the storytree is the beginning of the story. The earliest article of the root node is the root article. If the root article did not occur, the story might not exist. Therefore, the root article participates in multiple causal links. We extract the root article from each storytree and design an additional constraint, shown in Eq. 9,

$$\forall (i, j) \in Dp, x_{ij} \geq 1_{i \in R} \vee 1_{j \in R} \tag{9}$$

where R is the root article set of storytrees, the equation shows that the system fully encourages the root article to participate in causal relations. In [25], they defined the causal relation with 'the CAUSE event starts before the EFFECT event'. In our paper, the CAUSE event is the root article, and it is the earliest news article in the storytree, so the root article should have causal relations with all other news articles in its storytree. For ease of understanding, $R(i, j) = 1_{i \in R} \vee 1_{j \in R}$ is used below to indicate that at least one of i and j is the root node.

Different branch constrains (DFB) In our paper, we set up the storyline as a tree structure, each tree represents a whole story, and each branch represents a trend of the story. If two

articles are not on the same branch, they are not on the same trend of the story. Therefore, we design the constraint about it to discourage the causal relations between two articles in different branches in Eq. 10.

$$\forall(i, j) \in Dp, x_{ij} \leq 1_{R_i \subset R_j} \vee 1_{R_j \subset R_i} \tag{10}$$

In Eq. 10, R_i, R_j are the routes from the root node to node i and j, respectively . If $R_i \subset R_j$ or $R_j \subset R_i$, it means that i and j are in the same branch. We denote this by $B(i, j)$ below.

Same branch time constraints (SMBT) In the storytree, there is a problem with news articles on the same branch violating the definition of causality: The article near the root node may occur later than the article away from the root node. We set a constraint to discourage the causal relations of the news article pair that satisfies this condition. It is shown in Eqs. 11 and 12,

$$\forall(i, j) \in Dp, x_{ij} \leq \neg R(i, j)B(i, j)(|1 - L(i, j) - 1_{t_i < t_j}| + \frac{p_{ij}}{k_{sb}}) \tag{11}$$

$$L(i, j) = 1_{R_i \supset R_j} \wedge 1_{|R_i| < |R_j|} \tag{12}$$

where t_i and t_j are the time of the news article i and j, respectively. $|R_i|$ and $|R_j|$ are the length of R_i and R_j. $L(i, j)$ shows the relative position and time order of article i and j, $\neg R(i, j)B(i, j)$ means that news articles i and j are in the same branch and neither are in the root node, we increase p_{ij} by hyperparameter k_{sb} to reducing the impact of p_{ij} on ILP model.

Same branch M&S constraints (SMBM) Besides, we discourage causal links in the pairs not have root article. We refer to the first article at each node as the mainline article and the others as the side articles when two news articles are in the same branch. Therefore, four scenarios emerge from the combination of different types of news articles and location relations: MM (A mainline article is closer to the root node than the other), MS (A mainline article is closer to the root node than a side article), SM (A side article is closer to the root node than a mainline article) and SS (A side article is closer to the root node than the other). MM is more likely to have causal relations than the other three cases because there are mainline events on the same branch, so the constraint should be looser. The constraints are shown in the following equations:

$$\forall(i, j) \in Dp, x_{ij} \leq \neg R(i, j)B(i, j)(s_{mm} + s_{ms} + s_{sm} + s_{ss}) \tag{13}$$

$$s_{mm} = (1_{i \in M} \wedge 1_{j \in M}) \frac{p_{ij}}{k_{mm}} \tag{14}$$

$$s_{ms} = L(i, j)(1_{i \in M} \wedge 1_{j \in S}) \frac{p_{ij}}{k_{ms}} \tag{15}$$

$$s_{sm} = L(i, j)(1_{i \in S} \wedge 1_{j \in M}) \frac{p_{ij}}{k_{sm}} \tag{16}$$

$$s_{ss} = (1_{i \in S} \wedge 1_{j \in S}) \frac{p_{ij}}{k_{ss}} \tag{17}$$

where M and S are the set of mainline articles and side articles, respectively, $\neg R(i, j)B(i, j)$ represents that news articles i and j are not in the root node and in the same branch, k_{mm}, k_{sm}, k_{ms} and k_{ss} are hyperparameters. In the above equations, we can find that Eqs. 11 and 13 contain Eq. 10, but in later experiments, we will verify the validity of each type of constraint, so Eq. 9 is necessary. Moreover, two cases of this type of constraint will be discussed separately in Sect. 4, because although they belong to the same class, they describe different problems and may have different effects on the model.

Of the five constraints we designed, two encourage causal relation among news articles (SMN, RTN), and the remaining constraints (DFB, SMBT, SMBM) discourage causal relation among news articles. The SMN constraint describes two news articles about the same node, it does not conflict with the RTN which also encourages causality, the remaining constraints describe the relationship between different nodes in the story tree, and therefore do not conflict with the SMN constraint. The RTN constraint describes the case where there is at least one news article in the root node, while the DFB constraint describes two news articles in different branches, so there is no news article in the root node. In the formulas of the SMBT constraint and the SMBM constraint, $\neg R(i, j)$ indicates that these two constraints will not conflict with RTN constraints. Therefore, the five constraints in our model do not conflict in an integer linear programming model.

5 Experiments

There is no publicly available inter-document causal relation extraction dataset in the existing research that contains multiple article stories and labeled causal relations. Therefore, we constructed three datasets by crawling several events from 2018 to August 2020 from Yahoo, Reuters, and CNN. We save the title, text, time, and link information. These events spanned more than a week, with more than 5 reports. In fact, there are not many events with continuous news article coverage, so it is very difficult to manually discover many events that meet the above conditions. We use EDA [43] to augment data on these events with continuous news article coverage. EDA is state-of-the-art data augmentation technology for text classification tasks that consists of four basic operations: synonym replacement, random insertion, random swap, and random deletion. For a news article d, its keyword set $K_d = \text{En} \cup \text{Kw}$ is an important feature. According to the definition in Sect. 4.1, K_d is the keyword set of news article d, En is the set of named entities that appear more than once in d, and Kw represents the words with the highest TF-IDF. We first replaced the named entities in the news articles with other random names, so that the object described in the news article has changed, making the new news article d_e different from d. In order to maintain the association between the news articles, the same-named entities in different news articles are replaced by the same name. According to EDA, we replace the keywords in the article with synonyms, resulting in a new news article d_e^w, although d_e^w has the same structure as d, several news articles have completely different keywords. Random deletion is not utilized because this may delete some named entities and keywords. In addition, we collected some noise data D_n that are not related to existing news articles and that are not correlated with each other. We transform the dataset $D = \{d_1, d_2, ..., d_{|D|}\}$ into $D_{\text{aug}} = D \cup D_e^k \cup D_n$ through data augmentation and noisy data collection, where $D_e^k = \{d_{1e}^k, d_{2e}^k, ..., d_{|D|}^k\}$. In the end, we got three datasets with more than 1000 news articles. We labeled the news article pairs according to Wikipedia's definition of the causal relation. Table 1 shows the statistics of the datasets. There is no linked information in the Reuters dataset because the news on the Reuters website does not contain links.

In order to verify that the model is not affected by event types, we collected different categories of news. Table 2 shows that our event categories include politics, sport, social, and finance. Because there is no news labeled as politics in the search pages of Yahoo and CNN, we did not collect news on the topic of politics when forming these two datasets. Some events, such as the death of Kobe Bryant, belong to both sport and social categories. However, we only classify events into major categories to clearly show the distribution of

events. We randomly selected 20,000 article pairs for training. In order to ensure the number of positive examples, we randomly selected 90% of the positive examples, and the rest were supplemented with negative examples.

5.1 Baseline models

SVM Support vector machine (SVM) [9, 28] is a kernel method that finds the best separation hyperplane in the feature space to maximize the interval between positive and negative samples on the training set. The training data are mapped to a high-dimensional version through the so-called feature map to find the decision boundary.

GaussianNB GaussianNB [33] is one of the naive Bayes classifiers, which is a series of simple probabilistic classifiers based on the use of Bayes' theorem under the assumption of strong independence between features. GaussianNB deals with continuous data by assuming that the continuous values associated with each class are distributed according to a normal (or Gaussian) distribution.

AdaBoost Adaptive Boosting (AdaBoost) [11] is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. The AdaBoost method is an iterative algorithm that adds a new weak classifier in each round until it reaches a predetermined and sufficiently small error rate. In each iteration, AdaBoost can focus on samples that are more difficult to classify by increasing the weight of samples that are not correctly classified.

RandomForest RandomForest [4, 28] is a classifier that contains multiple decision trees, and its output category is determined by the mode of the output category of individual trees. Since

Table 1 Information of datasets from 3 source

Data source	#News article	#Word	#Sentence	#Relational fact
Yahoo	1046	333k	24,215	1003
Reuters	1212	197k	12,925	1596
CNN	1190	370k	26,277	1438

The number of relational fact is a two-way relation of calculation: when article pair (i, j) are counted, (j, i) are also counted

Table 2 The number of news articles of different categories

Dataset	Category			
	Politic	Sport	Social	Finance
Yahoo	–	250	506	290
Reuters	215	105	522	370
CNN	–	255	462	417

The major categories are based on how the news sites classify each article. Each story contains more than 5 articles. The stories include Trump's impeachment (politic), the death of Kobe Bryant (sport), Missing of a boy in Colorado (Social), etc.

each split is performed on a randomly selected subset of predictions, the tree is decorrelated, and the algorithm can improve the result by reducing the variance.

DER DER [2] is a neural model augmented by different grained text representations for implicit discourse relation recognition. The model consists of three parts: word-level module, sentence-level module, and pair-level module. The token sequences of sentence pairs are encoded by a word-level module first, and every token becomes a word embedding augmented by subword and ELMo. Then, the embeddings are fed to a sentence-level module and processed by CNN or RNN encoder blocks. In addition, the output of each layer is processed by the bidirectional attention module in the pair-level module and connected to the pair representation, and finally sent to the classifier. This model achieved state-of-the-art performance on PBTD 2.0.

We employ these baseline models on three datasets and take all the features in Sect. 4.4 as input for SVM, GaussianNB, AdaBoost, and RandomForest models. For the DER model, we take news article text as input. We get the results of the baseline model on the dataset and compare them with our model.

5.2 Experimental settings

In each dataset, there are 20000 news article pairs and less than 8% of these pairs have causal relations. We randomly select 60% of the news articles as the training set, 20% as the validation set, and the rest as the test set for all classifiers. The experiment was repeated 5 times with resampling each time; we report precision, recall, accuracy, and $F1$-score for inter-document causal relation extraction, which are obtained by calculating the average of five results.

The weighting parameters for constraints about storytree are hyperparameters, including k_b, k_{sn}, k_{sb}, k_{mm}, k_{ms}, k_{sm} and k_{ss} are between 0 and 1. In Sect. 4, we have shown that constraints do not conflict, so we find the optimal value for each k separately by grid search on each dataset. We set the SVM kernel to 'rbf' and the penalty coefficient is 0.8. We set the priors to None in GaussianNB. The base estimator of AdaBoost is DecisionTreeClassifier. The number of DecisionTree is 10 in RandomForestClassifier. In Sect. 4, we explained that constraints do not conflict with each other. Therefore, ablation study is performed by gradually increasing the number of constraints in the order of constraint presentation.

5.3 Performance comparison of different models

The last eight rows of Table 3 show the performance of our model and the result after adding each type of constraint gradually. **LR** is the base model of our system. The result shows the LR model has high recall, but suffers from low precision. The **LR + ILP (storytree)** model uses only all the constraints about the storytree (SMN, RTN, DFB, SMBT, SMBM) in the ILP model, basic constraints are not used in this model, and the purpose is to directly observe the impact of the constraints about storytree on the LR model. We can intuitively observe that the precision and $F1$-score of the storytree are improved on three datasets, especially the precision is increased by 31% to 42%. In the Reuters dataset, despite the loss of 4.5% recall, precision has been increased by 42.6%, resulting in a salient increase of over 19.7% in $F1$-score. The following results show how each constraint affects the performance of the model.

Table 3 Performance of different models on causal relation extraction

Models	Yahoo				Reuters				CNN			
	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc
SVM	78.0	78.5	78.2	97.8	75.6	80.3	77.8	96.6	79.0	82.3	80.6	97.4
GaussianNB	51.7	89.4	65.5	96.9	60.2	86.9	71.1	94.4	63.0	86.6	72.9	95.9
AdaBoost	72.6	80.0	76.1	97.5	66.9	80.5	73.0	96.5	75,7	83.2	79.3	97.2
RandomForest	78.3	81.7	79.9	98.1	75.8	80.4	78.1	96.7	79.9	81.6	80.8	97.4
DER [2]	62.3	71.8	66.9	97.0	59.4	70.4	64.4	92.4	65.2	72.6	68.7	95.1
LR	58.7	81.3	68.2	97.1	54.5	86.5	66.9	93.1	59.1	82.7	68.9	95.4
LR + ILP(storytree)	76.9	86.7	81.5	98.4	77.7	82.6	80.1	97.2	82.2	82.3	82.2	97.7
LR + ILP(basic)	57.6	**89.9**	70.2	97.2	55.9	85.1	67.5	93.3	59.1	86.2	70.1	95.0
+SMN	59.5	89.1	71.3	97.2	57.2	86.5	68.9	93.7	60.9	86.7	71.5	95.6
+SMN+RTN	62.0	88.8	73.1	97.3	57.4	**90.3**	70.2	94.0	60.7	**89.0**	72.1	95.8
+SMN + RTN + DFB	74.0	87.4	80.1	98.1	73.4	86.7	79.5	97.0	76.2	84.2	80.0	97.3
+SMN + RTN + DFB + SMBT	75.7	83.0	81.4	98.4	78.2	85.2	81.6	97.6	81.2	82.4	81.8	97.6
+SMN + RTN + DFB + SMBT + SMBM	**80.8**	88.6	**84.5**	**99.0**	**79.6**	86.0	**82.7**	**97.9**	**85.7**	84.7	**85.2**	**98.5**

LR + ILP (storytree) uses all constraints related to storytree (SMN + RTN + DFB + SMBT + SMBM) without using basic constraints. LR + ILP (basic) uses only basic constraints
Bold show the performance of our model as the best

LR + ILP (basic) model uses only the basic constraints in ILP, the result shows that this model has a subtle improvement in both $F1$-score, and the improvement in the recall is significant in Yahoo and CNN datasets; the recall in Yahoo reaches the highest value. However, it does not fundamentally solve the problem of the low precision of the model. **+Same Node Constraints (SMN)** shows the performance of the ILP system with constraints about encouraging the causal relation between the news articles from the same node in the storytree. In Reuters and CNN datasets, this constraint slightly increases recall and precision, resulting in a subtle improvement on the $F1$-score. In the Yahoo dataset, the addition of this constraint slightly reduces recall but still has an improvement on the $F1$-score. The results of **+Root Node Constraints (RTN)** are improved in different degrees for all three models, especially on the Yahoo dataset. On the CNN dataset, where the previous model does poorly, our model also gets some improvement. Although this constraint does not significantly improve the $F1$-score, the recall in the CNN and Reuters datasets has reached the highest value. Next, adding constraints about articles from different branches of storytree(**+Different Branch Constraints (DFB)**) improves the performance of the model. Although it reduces the recall on all three datasets, it significantly improves the precision and $F1$-score. **+Same Branch Time Constraints (SMBT)** reduces the recall of the three datasets, but also improves the accuracy rate again, making the precision and the recall closer, which improves the $F1$-score of all datasets. **+Same Branch M&S Constraints (SMBM)** improves precision and recall on the three datasets. The precision on the three datasets reaches the highest value and also has a high recall, which makes the $F1$-score of the model reach its peak.

In general, the ILP model based on storytree improves the performance of the model comprehensively, with over 30% improvement in all three datasets. Each constraint has a positive effect on the model. Different constraints have different effects on the model. Different Branch Constraints (DFB) have the most significant impact on the model.

The first five rows of Table 3 show the experiment results of other classifiers. SVM and RandomForest are the prevailing classifiers in inter-article relation classification, while

Table 4 Comparison of model performance before and after removal of link features

Models	Yahoo		CNN	
	F1 (with link)	F1 (without link)	F1 (with link)	F1 (without link)
SVM	78.2	77.9	80.6	78.9
GaussianNB	65.5	65.5	72.9	72.5
AdaBoost	76.1	73.3	79.3	77.1
RandomForest	79.9	78.5	80.8	78.8
LR + ILP (Full)	84.5	82.4	85.2	82.6

LR + ILP (Full) is the model with all the constraint

GaussianNB and AdaBoost are other text classifiers. The results show that AdaBoost and GaussianNB perform differently on the three datasets. AdaBoost performs the worst on the Reuters dataset, while GaussianNB performs the worst on the Yahoo dataset, and both perform best on the CNN dataset. RandomForest is the best-performing classifier. SVM's performance is slightly inferior to RandomForest. Both have higher accuracy and recall rates, but our model is higher in precision and recall. Compared with the best-performing classifier, our model has improved the $F1$-score by 5.7%, 5.9%, and 5.4%, respectively. In addition, our model also achieves the best performance on the accuracy metric.

Because the existing causal relation extraction models are not for inter-document, we use the model of discourse relation classification for comparison and causal relation is a type of implicit discourse relation. DER experimented on PBTD 2.0. We modified the preprocessing part of this model so that it can process our dataset. In their model, ELMo is used for text representation. When the length of the input text becomes longer, the training speed becomes slower. Each news article in our dataset is long. The average length of the CNN and Yahoo datasets exceeds 300 words, and the average length of the Reuters dataset is also close to 200 words, which causes the model to train very slowly and cannot converge. In order to solve this problem, we limited the length of the input text to 100 characters. Although the training speed is still very slow, the model finally converged and got the result. The result shows that our model overperforms this model by more than 24%.

To sum up, our storytree-based model not only improves the performance of the classifier but also achieves better results than the widely used classifiers and models. However, it is also clear that most of the methods perform worst on the Reuters dataset. There is no link information in the Reuters dataset, so we will verify how the link characteristics influence the model performance in the following experiment.

5.4 Validation of link feature

We assume that the reason the model does not perform as well on the Reuters dataset as the other two datasets is that it does not have linked information. In order to prove the validity of the link feature for the model, we remove the link information from the datasets of Yahoo and CNN and then train the full model and other classifiers in the same environment as before. The results are shown in Table 4.

The results show that the $F1$-scores of the models decrease on both datasets after the link feature is removed. GaussianNB is the least affected, with a slight decrease in the CNN dataset and the SVM is most affected. The $F1$ values of our model on these two datasets

Table 5 The number of misclassified classifiers under the same test set

Number of misclassified models			
1	2	3	4
(0,42), (10,6) (10,9), (0,2) (10,1)	(0,6), (1,6)	(0,9)	(10,8), (0,3)

The column of (a, b) indicates that n classifiers misclassify the relation between news article a and b

happen to be very close to the results on the Reuters dataset, and our model is still 5% higher than other models on three datasets. To sum up, the experiment results suggest that link information has a positive impact on the model, and the performance of the model on the three datasets is similar after the removal of link information.

5.5 Case study

In this section, we will use an example to explain how storytrees work in our model. We take the story of Kobe's death from the Yahoo dataset as an example and use the same test set to test our model and four other classification models. Then, we count the article pairs associated with the story in the results, which are misclassified by the classifier but correctly classified by our model. The results are shown in Table 5. Besides, the storytree containing this story is shown in Fig. 3. Each node represents an event. Inside the node are the news articles contained by the event. The root article is the report of Kobe's crash, and the others are all related after that. The dotted line linking the two articles represents the misclassified article pairs, which correspond to Table 5.

In this case, (10,8) and (0,3) are misclassified by all four classifiers. In article 10, former NBA star Wade expressed his condolences to Kobe Bryant in an interview a few days before the retirement ceremony (A.6). In article 8, Wade expressed his condolences to Bryant in a speech at the retirement ceremony (A.4). The two articles have a causal relation, but the classifiers do not classify them correctly. In the storytree, the two news articles are about the same event. In our model, the causal relation between articles in the same node is encouraged, and the article pair also meets the corresponding constraints, so our model correctly classified this article pair. For another example, article 0 is about Kobe Bryant's eldest daughter Natalia paying tribute to her late father and sister at the winter formal (A.1). This article is the latest in this event, and there are many descriptions of Natalia in this news article, which cannot be accurately classified only by similarity and keyword characteristics. But in the storytree, article number 3 is the root article (A.3), and our model encourages the root article to have causal relations with any articles in the same storytree, so our model is also classified correctly.

For (0,9), three classifiers determine causality, but article 9 is about Beyonce opening Kobe Bryant's memorial by singing Kobe's favorite song(A.5). There is no causal relation between the two articles. In the storytree, the node of article 0 is a child of the node of article 9. Our model also has constraints on events under the same branch. LR has low confidence scores on these two articles, so the final result of the ILP model is that there is no causal relation.

Article 42 is related to the new coronavirus and does not belong to this storytree(A.7), so it has no causal relation with article 0. SVM misclassifies the relation of the pair of articles, but our model does not encourage articles in different branches to have a causal relation.

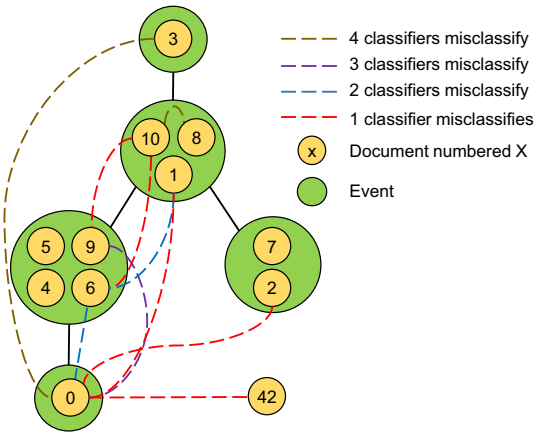


Fig. 3 A storytree of Kobe Bryant's death. The Numbers 0 to 10 represent 11 news articles about this event, and the number 42 is a article that is not related to this event

The same storytree also satisfies the condition, so our model judges wrong, (0, 2(A.2)) is the same.

From this case, it can be seen that each type of constraint can have a positive effect on the model. In general, using the storytree to design the constraints in the ILP model is effective in our model.

6 Discussion

In Sect. 5, we prove the effectiveness of the model by experiments, and it is better than other classification models. We conduct an ablation study by adding constraints gradually. From the results, we can know that the effect of each constraint on different datasets varies. Because the datasets are collected from different data sources, they are different in content, type, and language habits. After building a storytree, each storytree has a different structure, so the constraints play different roles. The most effective constraint is the Different Branch Constraint. We analyze the reason for this result is that the proportion of positive examples in the data is very small, and most article pairs are not in the same branch of the storytree. This constraint affects the most number of articles, leading to the most obvious improvement to the model. To prove our analysis, we extracted a small part of the dataset as a small version the dataset. The small version of the dataset only contains articles that can form a story and do not contain noise data. In each full dataset, there are more than 30 storytrees containing more than 5 news articles, while there are less than 5 storytrees in each small version dataset. We also conduct an ablation study on small datasets and compare it with the results of the ablation study on full datasets, as shown in Fig. 4. To eliminate the influence of constraints on the different strictness of the model, we set the constraint parameters the same as the full datasets.

As can be seen from the figure, for small datasets, Root Node Constraints have the largest impact on the Yahoo dataset, and Same Branch M&S Constraints have the largest improvement on the other two datasets. There are few storytrees in the small Yahoo dataset. The similarity between the root article and the later events is not high, resulting in errors in the LR classification results, so Different Branch Constrains affect more news articles. For the other two small datasets, the number of storytrees generated by the articles is deep, and there are few storytrees, resulting in many events in the same branch, so the constraints related to

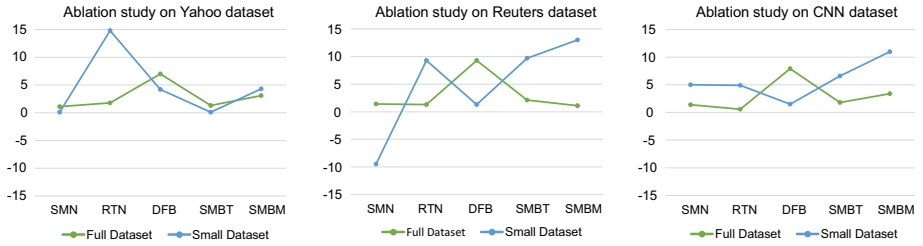


Fig. 4 Comparison of the results of ablation experiments on full datasets and small datasets. The horizontal axis is the constraint, and the vertical axis is the $\Delta F1$ of the model. A positive value represents an improvement, and a negative represents a decrease

branch (SMBT and SMBM) have a greater impact. In all the full datasets, Different Branch Constrains are the constraint condition for the highest improvement. This is also because most of the randomly selected news articles in each full dataset are not in the same storytree, which leads to the widest impact of Different Branch Constrains. In addition, we have also adjusted the parameters of Different Branch Constrains to make them less strict, and its impact is still the largest. Therefore, for our dataset, constraints that affect more news articles bring greater improvement.

In our model, the storytree affects the model in the form of constraints, but it means more to the model. In the RE task, people pay more attention to the relation between the sentence level and the document level, because in the same sentence or the same document, the entity pair is contextual, which can be used as a very important feature, and used in more complex models. But for the task of relation extraction of inter-documents, each news article is more independent, and there is no contextual information between articles. The storytree system clusters news articles into stories based on datasets and stores them in the form of trees. This is exactly the contextual information that has the time sequence and relative position for the news articles. This is a new method for feature-based models. Other feature-based models focus on improving the quality of features [10] or improving the classifier [13]. We not only use the basic information of the data to build features but build storytrees to find a deeper connection between news articles to create contextual information. The created contextual information is applied to the feature-based model, which is a method that the previous model did not propose. In addition, compared with the deep learning model, our model consumes far fewer resources and time than the deep learning model and obtains better results. In general, our model is different from the existing work, and it is also efficient and effective.

Although the ILP model improves both precision and recall, in the case study section, we found that some errors are caused by constraints. This is because the storytree does not completely represent the relative position between news articles, so the performance of the model depends on the accuracy of the storytree, which is a limitation of this method. In the future work, we hope to optimize the construction of the storytree or come up with new ways to construct contextual information between news articles.

The implication of our research is to propose a new method for extracting the causal relationship between long text data like news articles. Most of the previous models studied the extraction of relationships between entities, which is more microscopic than our research. However, there are not many studies on inter-document relation extraction. Our research provides an effective method, and we will improve our method in future work. On the other hand, the extraction of relations between long text data can have many applications. For example, the relation between scientific research papers can make people find more relevant

and logically continuous papers, and the relation between news can help people understand the entire event. By improving the accuracy of extracting relations between news articles, people can learn more about the causes and consequences of news when reading news.

7 Conclusion

In this paper, we propose a model for extracting causal relations of news articles based on storytrees. We illustrate the development of news articles by constructing storytrees for the data. Constraints are designed based on the structure information of the storytree, and the constraints are applied to the ILP system. The experimental results show that our model performs better on the three datasets than widely used classifiers and a state-of-the-art deep learning model. Our model is improved by 5.7%, 5.9%, and 5.4% compared to the best-performing classifiers. In future work, we will build a larger dataset, and optimize and improve the construction of the storytree.

Acknowledgements This work is supported by National Natural Science Foundation of China (Grant Nos. 71531001 and 61421003).

Appendix A. News articles in case study

In Section 5.6, we explain how story trees work in our model based on cases, and use numbers to represent news articles. In this section, we present the text of some of the articles mentioned in Section 5.6. For very long news articles, we only show a portion of the articles.

A.1. Article 0

Kobe Bryant's Daughter Natalia, 17, Pays Tribute to Late Dad and Sister Gianna at Winter Formal Kobe Bryant 's eldest daughter Natalia stopped to pose with a mural honoring her late dad and little sister Gianna as she headed to her winter formal. On Sunday, Vanessa Bryant shared a photograph of her 17-year-old all dressed up and ready to attend her high school dance. Ahead of the formal, Natalia posed for photographs in front of a tribute mural painted to honor Kobe, 41, and 13-year-old sister Gigi, both who died in the tragic Jan. 26 helicopter crash in Calabasas, California ."my babies. Natalia. #winterformal," the mom of four captioned the photograph, which featured a smiling Natalia, who was dressed in a blue and white polka dot dress. Fans flooded Bryant's comments section with compliments for Natalia. NBA star Dwyane Wade commented to heart emojis, while WNBA star Candace Parker wrote, "BEAUTIFUL" with several heart emojis. "She's beautiful and so is that mural. one fan wrote. Another added, "This warms my heart and at the same time saddens it. Good to see you girls pushing through. "RELATED: Vanessa Bryant 'Devastated' by Claims of L.A. Deputies Sharing Photos of Helicopter Crash Site The post comes just one week after Vanessa's legal team spoke out about the allegations that Los Angeles County Sheriff's deputies shared graphic photos of the helicopter crash site where Kobe, Gigi and seven others were killed on Jan. 26. In the statement, they denounced the "inexcusable" acts "of injustice" and called on "an Internal Affairs investigation of these alleged incidents. "The Los Angeles County Sheriff's office also released a statement, claiming an investigation surrounding the allegations was underway. Last month, Vanessa, 37, also filed a wrongful

death lawsuit against the helicopter company that owned the aircraft in the tragic crash. RELATED: Powerful Kobe & Gianna Bryant Fan Art Created to Honor Their Legacies In a complaint obtained by PEOPLE that lists herself and her daughters as plaintiffs, the NBA star's widow is suing Island Express Helicopters and claims that pilot Ara Zobayan of Huntington Beach, California, who was piloting the flight at the time of the crash and died , "failed to properly monitor and assess the weather prior to takeoff," "failed to abort the flight when he knew of the cloudy conditions" and "failed to properly and safely operate the helicopter resulting in a crash. "The complaint also claims that Island Express Helicopters "knew or should have known" that Zoboyan had been previously cited by the FAA for violating "the visual flight rules minimums by flying into an area of reduced visibility from weather conditions. "Vanessa and her daughters are seeking general, economic and punitive damages. In response to the lawsuit, a spokesperson for Island Express Helicopters told PEOPLE, "This was a tragic accident. We will have no comment on the pending litigation.

A.2. Article 2

KISS Pay Elaborate Tribute to Kobe Bryant at Staples Center: Watch The post KISS Pay Elaborate Tribute to Kobe Bryant at Staples Center: Watch appeared first on Consequence of Sound .KISS paid tribute to late NBA legend Kobe Bryant during their show at Staples Center in Los Angeles on Wednesday night. Bryant, 41, and his 13-year-old daughter, Gianna, died in a tragic helicopter crash along with seven others in late January, and KISS' Paul Stanley took a moment to show his respect. Donning Bryant's No. 24 Lakers jersey, Stanley gave a brief monologue during the band's set before playing Destroyer track "Do You Love Me". "We're in the house that Kobe built," Stanley said of the Lakers' home arena. "None of us would be here if this place wasn't really like a memorial to somebody who was so much more than a basketball player, somebody who's been a role model. And tonight, I think we dedicate this show not only to Kobe and his daughter Gigi, but to all the people who perished on that helicopter. "Bryant's retired numbers - No. 8 and 24 - were displayed on the band's stage screens as they performed "Do You Love Me". As the song concluded, Lakers-colored purple and gold balloons flooded the stage and Stanley dribbled them like basketballs. Editors' Picks Stanley previously expressed his sadness after Bryant's death on Twitter . Posted with a picture of Bryant and himself shaking hands courtside, Stanley wrote: "WOW! Kobe. Such A Shock. My Condolences To His Wife And Children. Very, very sad. #KobeBryant". KISS will continue the U.S. leg of their "End of the Road Tour" with Van Halen's David Lee Roth through mid-March before heading across Europe this summer. After that, they'll return to the States for another leg in the fall. Get tickets to KISS' upcoming shows here .Watch KISS' tribute to Kobe Bryant below. Popular PostsSubscribe to Consequence of Sound's email digest and get the latest breaking news in music, film, and television, tour updates, access to exclusive giveaways, and more straight to your inbox.

A.3. Article 3

Kobe Bryant's death leaves sports world stunned Kobe Bryant was killed in a helicopter crash in Calabasas Sunday morning, a source confirms to PEOPLE. The NBA legend, 41, was reportedly traveling with at least three other people in his private helicopter when it went down, according to TMZ. Emergency personnel responded but nobody on board survived. Five people are confirmed dead, TMZ reported. The outlet says that Bryant's wife, Vanessa Bryant, was not onboard. Spokespersons for LA county sheriff's office and LAPD did not

immediately respond to PEOPLE's request for comment. Bryant is survived by Vanessa, 37, and their four children together: daughters Natalia, 17, Gianna, 13, Bianka, 3, and son Capri, 7 months. Since the start of his basketball career, Bryant was one of the most accomplished men in the NBA, having played all 20 seasons with the Los Angeles Lakers. Until yesterday, he was the third-leading scorer in NBA history with 33,643 points but was surpassed LeBron James. James paid tribute to Bryant with special Nike shoes during the game against the Philadelphia 76ers.

A.4. Article 8

Dwyane Wade Reflects on Kobe Bryant's Wish to Inspire Others in Jersey Retirement Speech Whenever Kemba Walker looks down at the No. 8 on his jersey, he'll be reminded of the Mamba Mentality. While others in the NBA switched their jersey numbers in the wake of Kobe Bryant's tragic death, Walker instead decided to honor the Los Angeles Lakers legend by keeping his No. 8. The Boston Celtics guard spoke with ESPN's Rachel Nichols about what it means to wear that number going forward. N̈ow, that number means even more. So every time I step on the court, I just want to give 100 percent for him,Ẅalker told Nichols. T̈hat's my goal for the rest of the year and for the rest of my career. L̈IVE stream the Celtics all season. In the immediate aftermath of Bryant's passing, Walker considered a number change. Evidently, he decided the best way for him to honor Bryant was to continue wearing No. 8 while putting 100-percent effort into every game, just as Kobe would.Ï had a talk about it with some close people in my circle,Ẅalker said. Ï definitely thought about giving it up but then I thought, I think Kobe would want me and allow me to wear it. We want to keep his legacy going. I know of a few of us that's kept it. We're all just going to go out there and do what we can to play as hard as possible for Kobe.Ẅatch below: Not enough Kemba Walker in your life right now? He and I sat down to talk about what it was like replacing Kyrie, wearing No. 8 for Kobe, and what he thinks the Celtics ceiling is this postseason: pic.twitter.com/VyNBM24AZs- Rachel Nichols (Rachel__Nichols) February 27, 2020In 46 games this season, Walker has averaged 21.8 points, 4.1 rebounds, and 5.0 assists while earning his fourth All-Star selection. A knee injury has kept Walker out of commission for the last few games, but the Celtics are hopeful they'll have him back in the lineup sooner rather than later. Don't miss NBC Sports Boston's coverage of Rockets-Celtics, which begins Saturday at 7:30 p.m. with Celtics Pregame Live. You can also Kemba Walker explains keeping No. 8 to honor Kobe Bryant: 'We want to keep his legacy going'.

A.5. Article 9

Beyonce opens Kobe Bryant's memorial by singing 'XO,' one of Bryant's favorite songs The memorial to Kobe and Gianna Bryant began in an inspiring way Monday. The memorial opened with Beyonce, who told attendees, "I'm here because I love Kobe," before launching into one of Bryant's favorite songs. Beyonce then began singing "XO." Beyonce opens Kobe & Gianna's Celebration of Life with one of his favorite songs.(via SpectrumSN) February 24, 2020, Beyonce followed that up with "Halo." Beyonce performs Ḧaloät Kobe and Gianna's memorial.#KobeFarewell - Entertainment Tonight (etnow) February 24, 2020Beyonce was backed up by a chorus and an orchestra. After Beyonce opened the event, Vanessa Bryant eulogized Kobe and Gianna in a moving speech. Bryant and his 13-year-old daughter Gianna were among the nine people killed in a helicopter crash in January. Fans gathered to put together a make-shift memorial outside the Staples Center in the days after Bryant's death.

The Staples Center decided to host a memorial for Kobe, Gianna and the seven other victims of the crash. February 24 - or 224 - was chosen as the date of the memorial. Gianna Bryant wore No. 2. Kobe Bryant wore No. 24. More from Yahoo Sports: Eisenberg: How Kobe touched the lives of 10 everyday peopleIole: Wilder assistant was right to throw in towel vs. FuryKeyser: How do Astros fans feel about sign-stealing scandal? Bucks clinch playoff spot faster than anyone in at least 15 years.

A.6. Article 10

Dwyane Wade Says Friend Kobe Bryant Was 'in the Process of Building' His Next Legacy Before Death Dwyane Wade says Kobe Bryant was just beginning his second act ahead of his shocking death in a helicopter crash in January. Wade tells PEOPLE in an interview ahead of the release of his documentary, D. Wade: Life Unexpected, that Bryant's legacy is so much more than his illustrious career in the NBA. "His legacy is what he was in the process of building that we all got a chance to watch, right?" says Wade, 38. "We've seen what he did for basketball. We've seen that legacy. "Continues the former Miami Heat star, "But the legacy he was building outside of there was being there for the players, being a voice for the next generation. Working them out, being on the court with them, being there in his kids' lives, being a real all-star, superstar parent. Being an amazing husband. "Bryant, 41, was married to Vanessa Bryant, 37. The couple shares four daughters, including Gianna, 13, who was also killed in the crash. "And I think the one thing Kobe told us along the way is that no one is perfect in this, but at some point in his life - I said this recently - he mastered all of it," Wade tells PEOPLE. "He started mastering all of this. And then he showed us, too, that, 'Listen, we can do anything we want.' "RELATED: Dwyane Wade Had to Rank Himself and His Former Heat Teammates LeBron James and Chris Bosh on Wade calls his friend's legacy "so huge," adding, "and I think the thing that hurt more so than anything is that we all feel that we lost a loved one when Kobe passed". "And that's powerful - for someone that a lot of people haven't even met or didn't even know, still are mourning and trying to get over it, trying to move on with life," the retired athlete reflects. "That's when you know that you've built something, you've created something special. "RELATED: Dwyane Wade on How He and LeBron James Are Different as Basketball Dads: I 'Have More Self-Talk' Immediately after Bryant's death, Wade released a video of himself crying on Instagram , admitting, "Today is one of the saddest days in my lifetime". "It seems like a bad dream that you just wanna wake up from. It's a nightmare." D. Wade: Life Unexpected from ESPN Films and Imagine Documentaries and directed by Bob Metelus premieres Sunday at 9 p.m. EST on ESPN.

A.7. Article 42

Person in Washington State Is First in U.S. to Die From Coronavirus, Authorities Say A middle-aged patient in Washington state became the first person to die from the 2019 novel coronavirus inside the United States, officials said on Saturday as they announced additional cases, including a nursing home that could become the next hot zone. At least 69 people on American soil have had confirmed cases of the novel 2019 coronavirus, which is believed to have originated in a large seafood and live animal market in Wuhan, China, where it killed thousands before spreading to dozens of other countries. One American also died in China earlier this month. The U.S. outbreak seemed to reach a new stage over the weekend, with the number of confirmed patients who contracted it locally not from traveling abroad-creeping up.

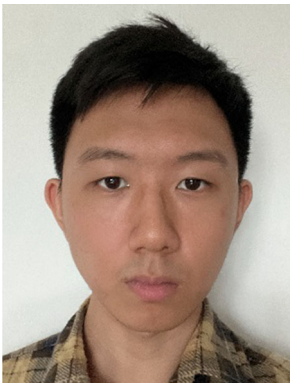
California announced Saturday that it had recorded a third such "community spread" case, a patient who was apparently infected by a Santa Clara County woman diagnosed a day earlier. The person who died in Washington state overnight was a man in his 50s considered at high risk, said Dr. Jeff Duchin, health officer for Seattle and King County. Dr. Robert Redfield, director of Centers for Disease Control and Prevention (CDC), said there was currently "no evidence" that the person who died had traveled recently to China or had any contact with someone who had-making it another case of "community spread" or unknown origin. "It's a tough one, but a lot of progress has been made," President Trump said at a press conference Saturday, stressing that the risk to the general population remained low. "We're doing really well," he added, "under incredibly adverse circumstances...

References

1. Aryal S, Ting KM, Washio T, Haffari G (2019) A new simple and effective measure for bag-of-word inter-document similarity measurement. CoRR, abs/1902.03402
2. Bai H, Zhao H (2018) Deep enhanced representation for implicit discourse relation recognition. In: Proceedings of the 27th international conference on computational linguistics, Santa Fe, New Mexico, USA, August 2018, pp 571–583
3. Blei D, Ng A, Jordan M (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
4. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
5. Cao J, Wang S, Wen D, Peng Z, Yu PS, Wang F (2020) Mutual clustering on comparative texts via heterogeneous information networks. Mach Learn 62:175–202
6. Cao Y, Fang M, Tao D (2019) BAG: bi-directional attention entity graph convolutional network for multi-hop reasoning question answerings. In: Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), Minneapolis, Minnesota, June 2019, pp 357–362
7. Chang M, Ratinov L, Roth D (2012) Structured learning with constrained conditional models. Mach Learn 88(3):399–431
8. Christopoulou F, Miwa M, Ananiadou S (2018) A walk-based model on entity graphs for relation extraction. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: short papers), Melbourne, Australia, July 2018, pp 81–88
9. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
10. El Barbary OG, Salama AS (2018) Feature selection for document classification based on topology. Egypt Inform J 19(2):129–132
11. Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139
12. Gao L, Choubey PK, Huang R (2019) Modeling document-level causal structures for event causal relation identification. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), Minneapolis, Minnesota, June 2019, pp 1808–1817
13. Goldberg AB, Zhu X, Wright S (2007) Dissimilarity in graph-based semi-supervised classification. In: Proceedings of the eleventh international conference on artificial intelligence and statistics), San Juan, Puerto Rico, March 2007, pp 155–162
14. Han X, Wang L (2020) A novel document-level relation extraction method based on Bert and entity information. IEEE Access 8:96912–96919
15. Haneczok J, Piskorski J (2020) Shallow and deep learning for event relatedness classification. Inf Process Manage 57:102371
16. Jiang H, Liu JT, Zhang S, Yang D, Xiao Y, Wang W (2020) Surface pattern-enhanced relation extraction with global constraints. Knowl Inf Syst 62:4509–4540
17. Sparck Jones K, Walker S, Robertson SE (2000) A probabilistic model of information retrieval: development and comparative experiments part 2. Inf Process Manag 36(6):809–40
18. Krishnamoorthy S (2018) Surface pattern-enhanced relation extraction with global constraints. Knowl Inf Syst 56:373–394
19. Liu B, Han FX, Niu D, Kong L, Lai K, Xu Y (2020) Story forest: extracting events and telling stories from breaking news. ACM Trans Knowl Discov Data 14(3):1–28

20. Liu P, Gulla JA, Zhang L (2018) Retracted article: a joint model for analyzing topic and sentiment dynamics from large-scale online news. World Wide Web 21:1527–1549
21. Lu T (2015) Semi-supervised microblog sentiment analysis using social relation and text similarity. In: 2015 international conference on big data and smart computing (BIGCOMP), February 2015, pp 194–201
22. Lv S, Huang L, Zang L, Zhou W, Han J, Songlin H (2020) RETRACTED ARTICLE: a joint model for analyzing topic and sentiment dynamics from large-scale online news. World Wide Web 23:2449–2470
23. Mele I, Bahrainian SA, Crestani F (2019) Event mining and timeliness analysis from heterogeneous news streams. Inf Process Manag 56(3):969–993
24. Morente-Molinera JA, Wikstrom R, Herrera-Viedma E, Carlsson C (2019) A linguistic mobile decision support system based on fuzzy ontology to facilitate knowledge mobilization. Decis Support Syst 81:66–75
25. Mostafazadeh N, Grealish A, Chambers N, Allen J, Vanderwende L (2016) CaTeRS: causal and temporal relation scheme for semantic annotation of event structures. In: Proceedings of the fourth workshop on events, San Diego, California, June 2016, pp 51–61
26. Nan G, Guo Z, Sekulic Ivan , Lu Wei (2020) Reasoning with Latent Structure Refinement for Document-Level Relation Extraction. Proceedings of the 58th annual meeting of the association for computational linguistics, Online, July 2020, pp 1546–1557
27. Ning Q, Feng Z, Wu H, Roth D (2018) Reasoning with latent structure refinement for document-level relation extraction. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers), Melbourne, Australia, July 2018, pp 2278–2288
28. Nordhausen K (2009) The elements of statistical learning: data mining, inference, and prediction. Int Stat Rev 77(3):482–482
29. Ohsawa Y, Benson NE, Yachida M (1998) KeyGraph: automatic indexing by co-occurrence graph based on building construction metaphor. In: Proceedings IEEE international forum on research and technology advances in digital libraries -ADL'98-, pp 12–18
30. Qin L, Zhang Z, Zhao H (2016) A stacking gated neural architecture for implicit discourse relation classification. In: Proceedings of the 2016 conference on empirical methods in natural language processing, Austin, Texas, November 2016, pp 2263–2270
31. Qin P, Xu W , Wang WY (2018) Robust distant supervision relation extraction via deep reinforcement learning. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers), Melbourne, Australia, July 2018, pp 2137–2147
32. Radinsky K, Horvitz E (2013) Mining the web to predict future events. In: Proceedings of the sixth ACM international conference on web search and data mining, New York, NY, USA, pp 255–264
33. Rish I (2001) An empirical study of the Naive Bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence, vol 3, pp 41–46
34. Robertson S, Zaragoza H (2009) The probabilistic relevance framework: BM25 and beyond. Found Trends Inf Retr 3(4):333–389
35. Roth B, Klakow D (2013) Feature-based models for improving the quality of noisy training data for relation extraction. In: Proceedings of the 22nd ACM international conference on information & knowledge management, New York, NY, USA, 2013, pp 1181–1184
36. Shahaf D, Yang J, Suen C, Jacobs J, Wang H, Leskovec J (2013) Information cartography: creating zoomable, large-scale maps of information. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining, New York, NY, USA, 2013, pp 1097–1105
37. Sheng Y, Zenglin X, Wang Y, de Melo G (2020) Multi-document semantic relation extraction for news analytics. World Wide Web 23:2043–2077
38. Sparck Jones K (1972) A statistical interpretation of term specificity and its application in retrieval. J Doc 28(1):11–21
39. Steen J, Markert K (2019) Abstractive timeline summarization. In: Proceedings of the 2nd workshop on new frontiers in summarization, Hong Kong, China, November 2019, pp 21–31
40. Tang H, Cao Y, Zhang Z, Cao J, Fang F, Wang S, Yin P (2020) HIN: hierarchical inference network for document-level relation extraction. In: Advances in knowledge discovery and data mining. Springer, Cham, pp 197–209
41. Vo D-T, Al-Obeidat F, Bagheri E (2020) HIN: hierarchical inference network for document-level relation extraction. Inf Process Manag 57(6):102319
42. Wang X, Jiang M (2020) Precise temporal slot filling via truth finding with data-driven commonsense. Knowl Inf Syst 62:4113–4139
43. Wei J, Zou K (2019) EDA: easy data augmentation techniques for boosting performance on text classification tasks. 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), Hong Kong, China, November 2019, pp 6382–6388

44. Changxing W, Chaowen H, Li R, Lin H, Jinsong S (2020) Hierarchical multitask learning with CRF for implicit discourse relation recognition. Knowl Based Syst 195:105637
45. Yang CC, Shi X, Wei C (2009) Discovering event evolution graphs from news corpora. IEEE Trans Syst Man Cybern Part A Syst Hum 39(4):850–863
46. Yao Y, Ye D, Li P, Han X, Lin Y, Liu Z, Liu Z, Huang L, Zhou J, Sun M (2019) DocRED: a large-scale document-level relation extraction dataset. In: Proceedings of the 57th annual meeting of the association for computational linguistics, Florence, Italy, July 2019, pp 764–777
47. Zhang F, Liu X, Tang J, Dong Y, Yao P, Zhang J, Gu X, Wang Y, Shao B, Li R, Wang K (2019) OAG: toward linking large-scale heterogeneous entity graphs. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, New York, NY, USA, 2019, pp 2585–2595



Chong Zhang received a B.E. degree from Beihang University, Beijing, China, in 2018. He is currently a Ph.D. student at the Department of Computer Science and Technology, Beihang University, Beijing, China. His research interests include natural language processing, graph learning and social media analysis.

Jiagao Lyu received a B.E. degree in 2017 and an M.E. degree in 2020 both from the School of Computer Science and Engineering, Beihang University, Beijing, China. His research interests include natural language processing and machine learning, focusing on information extraction and text analysis. He is also passionate about interdisciplinary topics such as bioinformatics and computational finance.

Ke Xu is a professor at Beihang University, China. He received his B.E., M.E. and Ph.D. degrees from Beihang University in 1993, 1996 and 2000, respectively. His research interests include algorithm and complexity, data mining and complex networks.