Neurocomputing 173 (2016) 1943–1950

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Event causality extraction based on connectives analysis

Sendong Zhao [a], Ting Liu [a,*], Sicheng Zhao [a], Yiheng Chen [a], Jian-Yun Nie [b]

[a] School of Computer Science and Technology, Harbin Institute of Technology, China
[b] Department of Computer Science, Université de Montréal, Canada

ARTICLE INFO

Article history:
Received 26 May 2015
Received in revised form
18 August 2015
Accepted 21 September 2015
Communicated by Rongrong Ji
Available online 1 October 2015

Keywords:
Causality extraction
Connective categorization
Hidden Naive Bayes
Text mining

ABSTRACT

Causality is an important type of relation which is crucial in numerous tasks, such as predicting future events, generating scenario, question answering, textual entailment and discourse comprehension. Therefore, causality extraction is a fundamental task in text mining. Many efforts have been dedicated to extracting causality from texts utilizing patterns, constraints and machine learning techniques. This paper presents a new Restricted Hidden Naive Bayes model to extract causality from texts. Besides some commonly used features, such as contextual features, syntactic features, position features, we also utilize a new category feature of causal connectives. This new feature is obtained from the tree kernel similarity of sentences containing connectives. In previous studies, the features have been usually assumed to be independent, which is not the case in reality. The advantage of our model lies in its ability to cope with partial interactions among features so as to avoid over-fitting problem on Hidden Naive Bayes model, especially the interaction between the connective category and the syntactic structure of sentences. Evaluation on a public dataset shows that our method goes beyond all the baselines.

1. Introduction

Considerable amounts of causality are available in natural language texts. For example, the sentence "Financial stress is one of the main causes of divorce" expresses a causal relation. Causalities extracted from texts can be useful in a number of tasks, such as event detection and prediction [1], generating future scenario [2,3] and question answering [4].

The present paper aims to mine causality as a kind of knowledge. Causality is usually evoked by verbal and non-verbal connective patterns (e.g., *cause*, *lead to*, *because of*) or lexico-syntactic patterns. In the above example, causality is evoked by the term *"causes of"*.

Numerous efforts have been dedicated to extracting causality from texts [4,2,5,6]. Similar to the extraction of other relations (e.g., *is-a* [7], *part-of* [8]), previous methods for causality extraction fall into two categories: methods based on patterns, including lexico-syntactic patterns, semantic relation patterns, self-constructed constraints and methods based on machine learning. Most of these approaches are heavily dependent on knowledge bases, such as Wikipedia, WordNet. However, they rarely consider the ways of causal connective evoking causality in the syntactic structure of the sentence, which is significant for causality extraction. As a result, they have poor performances on causality extraction without knowledge base.

In general, our work is based on two observations. On the one hand, some causal connectives show similarities in expressing causality, such as "increase" and "decrease". On the other hand, given a causal connective the sentence formation can partly be certain and given a syntax tree the pos of each node can be certain to some extent. Therefore, we use a new feature of connective category to encode the first observation and use the RHNB to model the second observation.

We can see from examples in Table 1, that a causal pair can be evoked by different causal connectives. In Examples 1 and 2, the causal pair is both (*Financial stress*, *divorce*). However, the connectives of these two causal pairs are different. Hence, the syntactic dependency structures of these two examples are different. The syntactic dependency structure of sentences expressing causality varies for different causal connectives. In other words, every causal connective has its own way to express causality. However, most of the causal connectives do not have a distinct way to express causality. Causal connectives are often used in similar ways to evoke causality, such as "increase" and "reduce". Therefore, we propose to divide causal connectives into different categories according to their ways to evoke causality. The goal of doing this is to fully use the collaborative filtering of instances

* Corresponding author.
E-mail addresses: sdzhao@ir.hit.edu.cn (S. Zhao), tliu@ir.hit.edu.cn (T. Liu), zsc@hit.edu.cn (S. Zhao), yhchenc@ir.hit.edu.cn (Y. Chen), nie@iro.umontreal.ca (J.-Y. Nie).

http://dx.doi.org/10.1016/j.neucom.2015.09.066
0925-2312/

Table 1
Examples of the same cause–effect pair which is evoked by different ways.

Example 1
Financial stress is one of the main [*causes of*] *divorce.*
$:\Longrightarrow(event_1$: Financial stress$)\overset{cause}{\longrightarrow}(event_2$: divorce$)$
Example 2
Financial stress [*increases*] *divorce.*
$:\Longrightarrow(event_1$: Financial stress$)\overset{increase}{\longrightarrow}(event_2$: divorce$)$

with similar causal connectives which may be in the same category.

In addition to categorization of causal connectives, there are many other features that may provide useful clues for causality extraction. We will include contextual features, syntactic features and position features. Based on those features, causality extraction can be cast as a classification among several candidates of causal pair. However, in previous studies, the features have been usually assumed to be independent, which is not the case in reality. For example, the categories of connectives and the syntactic features (see details later) may have strong interactions, i.e., they convey similar information. To cope with the possible interactions between features, we propose a new Restricted Hidden Naive Bayes (RHNB) model to extract causality, especially the RHNB model not only has good properties on learning hidden relations among features but also can avoid the over-fitting of relation learning on Hidden Naive Bayes.

In this paper, we develop a novel approach for causality extraction. The contributions can be summarized as follows:

- Causal connectives are divided into different classes by using the similarity of the syntactic dependency structure of sentences expressing causality, which significantly improves the causality extraction.
- We propose a RHNB model which is capable of coping with the interactions among features. In particular, the capability to cope with the interaction between causal connectives and lexico-syntactic patterns can highly improve causality extraction.

The remainder of the paper is organized as follows. We first discuss the related work. Then, expression of causality in English is discussed. And then, we introduce our new method for causality extraction, followed by the description of our experiments and results in detail. Finally, this paper ends with the conclusion and the future work.

2. Related work

Event causality extraction is a fundamental task because causality between event can be used in many applications. Causality is a strong principle to predict future. In order to predict future events, Ref. [1] extracted causal relations between events from a large-scale news corpus. Similarly, Ref. [3] proposed a supervised method of extracting event causality from the web to generate future scenarios. Causality is also an important resource for question answering. Ref. [4] tested on the question answering system with (61% precision) and without (36% precision) the causality module included. Ref. [9] explore the utility of intra- and inter-sentential causal relations between terms or clauses as evidence to answer why-questions better.

There are some approaches proposed by other authors concerning the automatic extraction and detection of causality. The main differences between our proposed method and the previous researches can be summarized as follows:

1. Our method uses a new connective category feature which is obtained from the tree kernel similarity of sentences.
2. Our method encodes the partial influence among features. For example, causal connectives usually determine the syntactic structures of the sentence and the relative position of cause and effect in a sentence.

For event causality extraction and detection, clues used by previous methods can be roughly categorized as lexico-syntactic patterns [10,1], words in context [9], associations among words [2,11] and the semantics of predicates and nouns [12,13]. Besides features similar to those described above, we propose a new connective category feature which is obtained from the tree kernel similarity of sentences containing causal connectives.

All the above studies except [4], only put emphasis on *Precision* rather than *Recall* or *F-sore* of their performance. In other words, such causality pairs cannot be acquired in any single sentence in their corpus with reasonable precision. There are some studies which aim not only the *Precision* but also the *Recall*, such as [4,14,6,5].

The models usually used in these studies on causality extraction, which pay attention to *F-sore*, are SVM, decision tree and some other semi-supervised methods. However, all these methods used in previous researches rarely consider the influence of interactions among features on causality extraction, especially the interaction between connective features and lexico-syntactic features, which is confirmed to be very effective by our experiments.

There are many studies on relation extraction [15–17] and discourse relation recognition [18–20]. The studies on relation extraction usually focused on extracting relations between named entities, which is different from causality extraction. Causal relation usually refers to the relationship between events.

3. Expression of causality

Numerous studies by linguists and philosophers over the last three decades have focused on causality, especially in terms of the semantics and syntax of causal constructions. This section introduces the major approaches in which causality is expressed in English. Here, we use the categories of causality proposed by [21]. Most of the time, causality is evoked by explicit causal connectives. Ref. [21] defined eight categories of explicit causal connectives and analyzed their frequency in the LOB corpus which was compiled by researchers at Lancaster University. The details are shown in Table 2.

Causality expressed by explicit connective refers to a causal situation where the causal relation is evoked by causal connectives such as “cause”, “effect” and “because of”. The sentence of Example 2, “Financial stress” is the nominal subject of the connective “increases” and “divorce” is the direct object of the connective. In this example, we can see that there is strong syntactic dependency behind the expression of causality. The cause event “Financial stress” and the

Table 2
Frequency of eight categories of causal connectives.

Categories	Types	% of total instances
Conjunction	12	28.7
Adverbs	9	28.0
Noun phrases	16	16.0
Complex prepositions	32	10.8
Explicit verbs	34	8.5
Verb phrases	23	5.0
Prepositions	3	2.8
Adjective phrases	3	0.2

effect event "divorce" have a same father node which is the connective "increases". The relations between father and children are "nsubj" (nominal subject) and "dobj" (direct object) respectively. This example may suggest that causality is usually expressed in some typical syntactic patterns. Indeed, a number of previous studies relied heavily on such patterns to determine causality. For example, Refs. [5] and [6] both used this kind of syntactic structure to detect causality. They perform well if causality is expressed in typical structures, such as "(.*) *increase* (.*)", "(.*) *cause* (.*)".

However, causality can also be expressed in other less typical structures. For example, the two sentences in Figs. 1 and 2 share the common syntactic structure "(.*) *from* (.*) *is increasing*". However, their semantic relations are different. The approaches of [5] and citeIttoo:2013 will fail to distinguish the cases. To solve this problem, we use additional contextual features. Contextual features include the causal words. In the above examples, we have, among others, "risk of" and "number of". Intuitively, the first group of words may hint a causality in the sentence than the latter. Several other types of features will also be included.

When many features are used, it is often the case that they interact. To cope with this interaction, we use the RHNB model for learning relations between contextual feature [22] and connective categories.

4. Our method

The input of our work is several sentences without annotation, and the output is the pairs of phrase which is decided as causality. The process of causality extraction is to extract some candidate pairs for each sentence and then determine which one is the causality. In this study, causality extraction is cast into a binary classification [23] framework.

In this section, we elaborate on our method to extract causal pairs from each sentence using RHNB model. We treat causality extraction as a classification problem. Therefore, we should use a machine learning model based on some useful features to extract causal pairs. On the one hand, we propose a new useful category feature of causal connectives besides some commonly used features. The new category feature of causal connectives is obtained from the similarity of the syntactic dependency structure of sentences expressing causality. On the other hand, we propose a new RHNB model to cope with the interactions among features, especially the interaction between causal connectives and lexico-syntactic patterns of sentences. This is a key factor that affects the performance of our method. The following parts will shed light on these two advantages of our method.

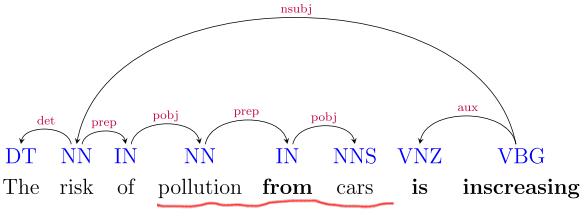


Fig. 1. POS tagging and dependency structure of a sentence with a causal relation. The sentence pattern is "(.*) *from* (.*) *is increasing*".

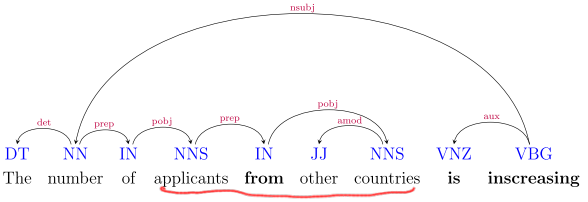


Fig. 2. POS tagging and dependency structure of a sentence without a causal relation. The sentence pattern is also "(.*) *from* (.*) *is increasing*".

In this work, we use some features to train the RHNB model. Table 3 illustrates the detailed features used to train RHNB. We select a list of 26 feature templates which are divided into four categories: contextual, syntactic, position and connective features.

- The contextual features are used to discriminate different semantic relations of pairs which have the same or similar syntactic structure, just like the situation of Figs. 1 and 2. Ref. [11] also indicated that more contextual knowledge seems necessary for better predictions of causal relation.
- Syntactic features reveal the syntactic relations between different parts of a sentence, which are important for identifying which parts of a sentence represent events of causal pairs. Ref. [24] demonstrated that syntactic features can improve to disambiguate explicit discourse connectives in text.
- Position features reflect the position of connective and the distance between the event and the connective. This kind of features is generally used in many other studies [14].
- The connective features are the features drawn from connective analysis. We find that some causal connectives have similar ways to evoke causality, which is meaningful for causality extraction. Therefore, we propose to divide causal connectives into different categories according to their ways to evoke causality. The goal of doing this is to fully use the collaborative filtering of instances with similar causal connectives which may be in the same category.

4.1. Causal connective categorization

In the section, we divide causal connectives into different classes according to the dependency structure of sentences involving these causal connectives. Causal connectives are often used in similar ways to evoke causality. Therefore, we propose to divide causal connectives into different categories. The goal of doing this is to fully use the collaborative filtering of instances with causal connectives in the same category. For example, an instance I_1 with connective c_1 in the training set and another instance I_2 with connective c_2 in the testing set. I_1 can be used to supervise the causality extraction from I_2 if c_1 and c_2 are in the same category. We use convolution tree kernels method [25] to calculate the similarity between two sentences involving causal connectives and then utilize this kind of similarity to divide causal connectives into different classes. Here, the tree kernel method is used to measure the similarity of syntactic structure of two sentences. The idea of tree kernel method is to count the number of the same sub-structures of the two parse trees. We define the similarity of two causal connectives as the mean similarity of two sets of sentences, which involve the two causal connectives. The

Table 3
List of features used to extract causality.

Contextual features	
$prev_1$, $prev_2$ of e	$prev_1$, $prev_2$ POS of e
$next_1$, $next_2$ of e	$next_1$, $next_2$ POS of e
$prev_1$, $prev_2$ of C	$prev_1$, $prev_2$ POS of C
$next_1$, $next_2$ of C	$next_1$, $next_2$ POS of C
Syntactic features	
The relation between e and its father	POS of its father
Syntactic relation between e and C	Length($C \longrightarrow e$)
Position features	
Dist(e,C)	Dist(Begin,C)
Causal connective features	
Include a C	C POS
Categories from Table 2	Clusters generated using tree kernel

formal definition is as follows:

$$Sim(c^{(i)}, c^{(j)}) = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N} K(T^{(m)}, T^{(n)})}{M \times N} \quad (1)$$

where $c^{(i)}$ and $c^{(j)}$ represent the ith causal connective and the jth causal connective. $K(T^{(m)}, T^{(n)})$ is the kernel function to measure the similarity degrees of the dependency trees of sentences in terms of the same sub-trees:

$$K(T^{(m)}, T^{(n)}) = \frac{|T^{(m)}_{sub} \bigcap T^{(n)}_{sub}|}{|T^{(m)}_{sub} \bigcup T^{(n)}_{sub}|} \quad (2)$$

where $T^{(m)}_{sub}$ is the set of sub-trees of tree $T^{(m)}$ and $T^{(n)}_{sub}$ is the set of sub-trees of tree $T^{(n)}$.

In order to divide causal connectives of our training data into different classes, a two-level K-means-hierarchical hybrid clustering method [26] is used according to the similarity of connectives $Sim(c^{(i)}, c^{(j)})$. An advantage of the two-level K-means-hierarchical hybrid clustering method is that the number of clusters is automatically determined. Then we use the category of causal connective as a feature shown in Table 3.

4.2. Restricted Hidden Naive Bayes

Since the intention of emphasizing on the relations among features, especially the relation between causal connective and syntactic structure of sentences, we use a Bayesian model in this section to extract causality from texts. Specifically, we propose Restricted Hidden Naive Bayes which is an improvement of Hidden Naive Bayes (HNB) [27]. RHNB inherits the structural simplicity of Naive Bayes and can be easily learned without structure learning.

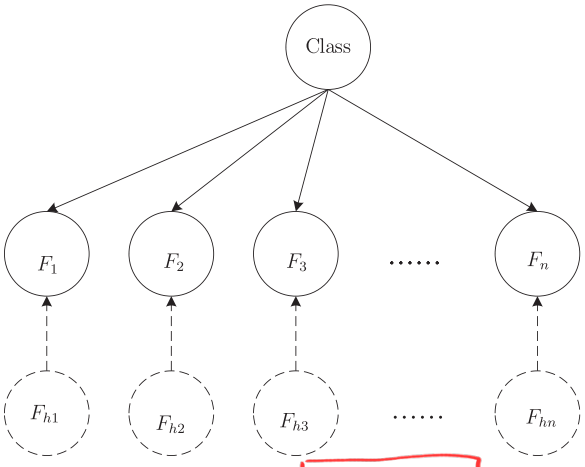


Fig. 3. The structure of Hidden Naive Bayes.

The basic idea of HNB is to create a hidden parent for each feature, which combines the influences from all the other features except itself. Fig. 3 gives the structure of an HNB. In Fig. 3, *Class* is the class node, and is also the parent of all feature nodes. Each feature F_i has a hidden parent F_{hi}, $i = 1, 2, \ldots, n$, represented by a dashed circle. The edge from the hidden parent F_{hi} to F_i is also represented by a dashed directed line, to distinguish it from regular edges. The edge of F_{hi} to F_i encodes the influences from all the features except F_i.

HNB creates a hidden parent for each feature, which combines the relations from all other features. In other words, HNB assumes that each feature can be affected by all the other features and learns all the influences from the other features. Actually, not all the features have mutual effects. Therefore, the assumption that all the features have mutual effects can easily lead to over-fitting problem with a small size of training dataset [28]. Usually, this kind of relation learning could learn some error relations or noisy relations along with correct relations [29–31]. This can affect the performance on the test set. Therefore, we propose the RHNB to solve the over-fitting problem of HNB on relations learning by restricting relations among the features. The advantage of our proposed RHNB model is that we suppose that part of the features have mutual effect but not all the features. This assumption can avoid learning some interactions between features which are error or noisy.

Fig. 4 gives the structure of the RHNB of causality extraction. In Fig. 4, *Class* is the class node, and is also the parent of all feature nodes. Features from F_1 to F_{26} are divided into four groups which are contextual, syntactic, position and connective. Each feature F_i has the same hidden parent F_{sub}, represented by a dashed circle. The arc from the hidden parent F_{sub} to F_i is also represented by a dashed directed line, to distinguish it from regular arcs. F_{sub} is the subset of feature templates set $(F_1, \ldots, F_{26})$. F_{sub} can be any combination of contextual features, syntactic features, position features and connective features. If we take F_{sub} as the whole feature templates set $(F_1, \ldots, F_{26})$. The RHNB model will degenerate into HNB model. The effect of F_{sub} on F_i is defined as follows:

$$P(F_i | F_{sub}, C) = \sum_{j=1, j\neq i}^{k} W_{ij} * P(F_i | F^{(j)}_{sub}, C) \quad (3)$$

where C denotes the class of instances, W_{ij} is the weight of edge from $F^{(j)}_{sub}$ to F_i (with $\sum_{j=1}^{k} W_{ij} = 1$). W_{ij} is defined as follows:

$$W_{ij} = \frac{I_P(F_i; F^{(j)}_{sub} | C)}{\sum_{j=1, j\neq i}^{k} I_P(F_i; F^{(j)}_{sub} | C)} \quad (4)$$

where the value of W_{ij} is the real impact of each connecting from hidden node F_{sub}. If W_{ij} is 0 then the impact is 0. $I_P(F_i; F^{(j)}_{sub} | C)$ is the

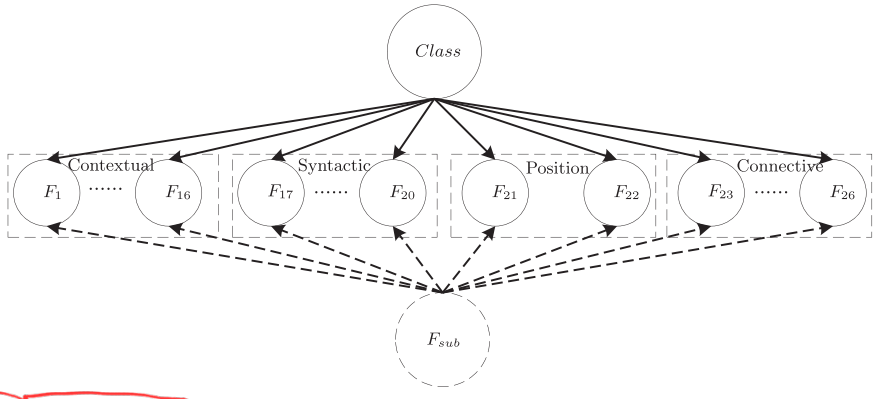


Fig. 4. The structure of Restricted Hidden Naive Bayes for causality extraction. F_{sub} is a hidden node which can be any subset of $(F_1, \ldots, F_{26})$. The hidden node has a direct effect on F_i except itself, i.e., if we choose the hidden node as (F_1, F_2), then (F_1, F_2) has effect on the left nodes $(F_3, \ldots, F_{26})$.

conditional mutual information, proposed by [32]:

$$I_P(F_i; F_{sub}^{(j)} | C) = \sum_{F_i, F_{sub}, C} P(F_i, F_{sub}^{(j)}, C) \times \log \frac{P(F_i, F_{sub}^{(j)} | C)}{P(F_i | C) P(F_{sub}^{(j)} | C)} \quad (5)$$

We take advantage of RHNB and features shown in Table 3 to extract causality from texts. The classifier RHNB is applied on an instance $I = (f_1, \ldots, f_{26})$ as follows to determine if the event pair is a causality:

$$Class(I) = \arg\max_C P(C) \prod_{i=1}^{14} P(F_i | F_{sub}, C) \quad (6)$$

The structure of RHNB is determined. Therefore, learning a RHNB is quite simple and mainly requires estimating the parameters in the RHNB from the training data. The parameters learning algorithm for RHNB is depicted as Algorithm 1.

Algorithm 1. Parameters learning algorithm for RHNB.

Input: A set D of training instances;
Output: A Restricted Hidden Naive Bayes for D;
  for each value C of class **do**
    Compute $P(C)$ from D
  end for
  for each feature F_i and $F_{sub}^{(j)}$ **do**
    Compute $P(F_i | C) P(F_{sub}^{(j)} | C)$ from D
  end for
  for each feature F_i and $F_{sub}^{(j)}$ **do**
    Compute $I_P(F_i; F_{sub}^{(j)} | C)$
  end for
  for each edge from $F_{sub}^{(j)}$ to F_i, which $F_{sub}^{(j)}$ and F_i is different features **do**
    Compute $W_{ij} = I_P(F_i; F_{sub}^{(j)} | C) / \sum_{j=1, j \neq i}^{k} I_P(F_i; F_{sub}^{(j)} | C)$
  end for

There are four steps in parameters learning algorithm for RHNB. Firstly, we should compute $P(C)$ from the training set D for each value C of class. Then, we should compute $P(F_i | C) P(F_{sub}^{(j)} | C)$ from the training set D for each feature F_i and $F_{sub}^{(j)}$. And then, we should Compute $I_P(F_i; F_{sub}^{(j)} | C)$ for each feature F_i and $F_{sub}^{(j)}$. Finally, we should compute W_{ij} for each edge from $F_{sub}^{(j)}$ to F_i. We can obtain all the parameters in the RHNB model after this process.

5. Experimental evaluation

Extensive experiments are conducted to test the effectiveness of the proposed method. In this section, we describe the experimental design and report the results to demonstrate the advantages of our method for extracting causality.

5.1. Data set

We use a corpus, which is composed of 2682 sentences, half of which contain causal relation. To evaluate, we split all the 2682 sentences into a training/validation/test set randomly, with the ratio of 8:1:1. The first is used for model training, the second for hyperparameter tuning on baseline methods likes SVM, and the third for evaluation. These corpora are obtained by extending the annotations of the SemEval-2010-Task8 dataset. In the original dataset, only one or two words have been annotated as the cause or effect in each sentence. We extended the cause and effect words as phrases. For example the annotations in the original dataset "The < e1 > *burst* < /e1 > has been caused by water hammer < e2 > *pressure* < /e2 >" modified to "The < e1 > *burst* < /e1 > has been caused by < e2 > *water hammer pressure* < /e2 >", where the effect is replaced by a more complete semantic unit *water hammer pressure*.

5.2. Preprocessing

In the process of dealing with sentences, we apply lemmatization and use mate tools to perform part-of-speech tagging and dependency parsing [33]. In most cases, the cause and the effect of causal pairs are not an individual word but a phrase. For example, "*heart and lung disease*" is an effect of "*smoking*", which can be extracted from "*Smoking can cause many types of heart and lung disease*". We also find that these kinds of phrases appearing as cause or effect are always in the form of noun phrases (NPs). Hence, in order to detect these NPs as semantic units which may be the cause and effect events, we run a partial parser [34] to identify NPs in sentences as a candidate.

5.3. Baselines

We used [6] and [5] methods as our baselines. Ref. [6] extracted causal pairs using the lexico-syntactic patterns between event pairs. These patterns are obtained from sentences on Wikipedia using a bootstrapping method by taking some causal pairs as seeds. We adapted these patterns to our training data. Ref. [5] just use an SVM classifier on some lexico-syntactic features and self-constructed constraints to extract causal pairs on a public dataset. They tested their method on the same dataset as ours, but they identified unigrams or bigrams as the cause and effect events, which is different from our method. For comparison, we added our phrase extraction module to their method.

Since we use the SemEval-2010-Task8 dataset, we want to compare our results with those obtained by the best system at SemEval. The UTD system [14] obtained the best performance of SemEval-2010. However, rather than extracting causal pairs of sentences, UTD system just predict if the unique labeled word pair of each sentence is a causal pair. That means for each sentence there is a unique labeled pair to be predicted for the UTD system. In the reality of causality extraction, there are more than one candidate pair of event in each sentence. Moreover, the UTD system used several external knowledge bases for relation classification, such as WordNet, NomLex-Plus, VerbNet, and the Google N-Gram data. Meanwhile, the method of UTD is not addressed in detail. Therefore, we think it is not proper for us to take UTD as our baseline.

5.4. Implementation details

We did not tune the K value. We use a two-level K-means hierarchical hybrid clustering method in which the K is determined automatically. A RBF kernel is used in SVM. We perform a grid search on the validation set for the best hyper-parameters C and γ and use them to get the result on test set.

5.5. Results and analysis

We conducted several kinds of comparison on causality extraction from texts. Generally, there are three types of comparison. The one is the comparison between our method and previous methods on the same dataset which is shown in Table 4. The one is the comparisons between classifiers using our features on the same dataset, as shown in Table 5. The other is the comparisons of different features and different usages of features, as shown in Tables 6–8 respectively.

How about computation cost ?

Table 4
Comparison in *P*, *R* and *F* on testing set with baselines.

Methods	P(%)	R(%)	F(%)
Ittoo2013	65.7	68.3	66.9
Sorgente2013	76.7	71.4	73.9
Ours	87.3	84.1	85.6

Table 5
Comparison of different classifiers on extracting causality.

Classifiers	P(%)	R(%)	F(%)
Naive Bayes	75.7	75.6	75.6
SVM (RBF)	80.1	81.2	79.6
Random Forest	83.6	83.5	83.5
HNB	86.1	82.3	84.1
RHNB	**87.3**	**84.1**	**85.6**

5.5.1. Results

Table 4 shows that our method outperforms the previous methods on all the three metrics. We conducted a significance test (*t*-test) on the improvements by our approach. The result indicates that the improvements are statistically significant (p-value < 0.05). In the one hand, our method uses more meaningful features in comparison with [5] and [6]. Ref. [5] just used lexico-syntactic features and some self-constructed rules. Ref. [6] just encode lexico-syntactic patterns in their bootstrapping framework. One the other hand, our RHNB model has more powerful ability to cope with the influence among the features than SVM used in [5]. The consideration of the interactions among the features, especially the interaction between contextual, syntactic feature and position and connective features, is an advantage of our method. Specifically, the syntactic dependency structure of a sentence expressing causality varies for different connectives. For example, the categories of causal connective affect the dependency structure of sentences. Usually, sentences describing the same causality with different connectives may require different structures. The examples shown in Table 1 illustrate this phenomenon.

5.5.2. Analysis

Based on the features in Table 3, we train several models based on Naive Bayes, SVM, Random Forest, HNB and RHNB on the training set. We can observe the performance comparisons among different classifiers on extracting causality from Table 5. It can be concluded from Table 5 that the more we consider the relation among features, the better performance we can obtain. However, when we suppose each feature can be affected by all the other features (i.e., the assumption of HNB) the performance will be worse than the assumption that each feature can be affected by the part of the other features (i.e., the assumption of RHNB). In this sense, the assumption of partial relations among features is better than assumption of complete relations among feature significantly (p-value < 0.05). In other words, the comparison between HNB and RHNB(F_{sub}=Syntactic+Connective) shows that RHNB can avoid over-fitting problem of HNB on relation learning.

In Table 6, we conducted six experiments by taking the F_{sub} as *NULL*, *Contextual*, *Syntactic*, *Position*, *Connective* and *Syntactic+Connective* respectively. RHNB(F_{sub}=NULL) means that the hidden node of RHNB is null. If the F_{sub} is null, the RHNB model degenerates into Naive Bayes. RHNB(F_{sub}=Contextual) means that we just select *Contextual* (i.e., $F_{sub}=\{F_1,\ldots,F_{16}\}$) as the hidden node of RHNB. Similarly, RHNB(F_{sub}=Syntactic+Category) means that we take the union of *Syntactic* and *Connective* as the hidden node of RHNB (i.e., $F_{sub}=\{F_{17},\ldots,F_{20}\}\cap\{F_{23},\ldots,F_{26}\}c$). We can see the contributions of our proposed *Connective* on causality extraction in

Table 6
Comparison of different hidden nodes F_{sub} of RHNB model on extracting causality.

Classifiers	P(%)	R(%)	F(%)
RHNB(F_{sub}=NULL)	75.7	75.6	75.6
RHNB(F_{sub}=Contextual)	75.7	75.6	75.6
RHNB(F_{sub}=Position)	75.7	75.6	75.6
RHNB(F_{sub}=Syntactic)	83.1	82.6	82.8
RHNB(F_{sub}=Connective)	84.4	83.1	83.7
RHNB(F_{sub}=Syntactic+Connective)	**87.3**	**84.1**	**85.6**

Table 7
The difference of decrease on P.R.F by deleting "Fang's Category" and "Tree Kernel based Category" respectively from F_{sub}=Syntactic+Connective.

F_{sub}=Syntactic+Connective	P(%)	R(%)	F(%)
Fang's Category	−0	−0.4	−0.2
Tree Kernel based Category	−3.6	−1.5	−2.5

Table 8
The influences of different kinds of features on causality extraction.

Methods	P(%)	R(%)	F(%)
Naive Bayes with ALL	75.7	75.6	75.6
Naive Bayes-Position	75.1	74.9	75.0
Naive Bayes-Contextual	73.8	66.6	69.1
Naive Bayes-Syntactic	64.2	61.7	62.9
Naive Bayes-Connective	56.2	52.8	54.4

RHNB by comparing RHNB(F_{sub}=Connective) with RHNB (F_{sub}=NULL). By the same way, we can also discern the contribution of *Contextual*, *Syntactic* and *Position* on causality extraction in RHNB in comparison with RHNB(F_{sub}=NULL).

Furthermore, we can see from the results that when the set F_{sub} as the union of syntactic features and connective features, the RHNB get the best result. The comparisons indicate the best *F* score in RHNB (F_{sub}=Syntactic+Connective). Syntactic feature is very useful to locate cause event and effect event. Specifically, different syntactic structures often has impact on POS tagging, position and connectives. Different connectives often generate different syntactic structures of sentences. For example, the position of causal events in the syntactic tree of sentence triggered by "lead to" is usually different from the syntactic tree of sentence triggered by "since".

There are two kinds of category feature shown in Table 3. In order to compare the effectiveness of these two kinds of category, we conducted experiments which are shown in Table 7. In order to test the contribution of "Fang's Category" and "Tree Kernel based Category" on event extraction, we conduct the following sanity check: (1) take F_{sub} as Syntactic+Category and get the result; (2) delete "Fang's Category" from F_{sub} and record the decrease on *P*, *R* and *F*; (3) delete the "Tree Kernel based Category" from F_{sub} and record the decrease on *P*, *R* and *F*; finally, we get Table 7.

Although we can see the contribution of each kind of features on causality extraction from the comparison in Table 6, the contribution is not pure. That is because the contribution also contains the effect of interaction between features. Therefore, in order to evaluate the importance of every kind of features, we also compare Naive Bayes classifier by dropping contextual features, connective features, syntactic features and position features respectively. Compared with other models, Naive Bayes classifier does not consider relation among features. Thus, it is a good model to assess the importance of the four kinds of features in Table 3. "Naive Bayes with ALL" in Table 8 means that we implement Naive Bayes model using all the four kinds of features. It is clear from the

comparison in Table 8 that the four kinds of features are all important for causality extraction. However, the importance to affect the result is different. From the comparison, the position features are the most important. If we do not use connective features the results are slightly better than random guessing. Syntactic features are the second most important. These two kinds of features had been largely discussed by the previous studies. It will lower the *F* score by 6.5% if we do not use contextual features. It will lower the *F* score by 0.6% if we do not use position features. Therefore, contextual features and position features are also useful to improve the performance of causality extraction.

6. Conclusion

In this paper, we proposed a RHNB model to extract causality from texts by using some useful features including contextual features, syntactic features, position features and connective features. On the one hand, we used some new features to extract causality. Especially, we divided causal connective into different categories by using the similarity of the syntactic dependency structure of sentences expressing causality. Categories of connectives learned by tree kernel method and proposed by [21] are used for the first time to improve the performance of causality extraction. On the other hand, we proposed a new model (RHNB) based on HNB, which not only inherit the relation learning ability among features but also avoid over-fitting problem on relation learning among features of HNB. Especially, RHNB has ability to handle with the interaction between causal connectives and lexico-syntactic patterns, which is proved to be an effective cue to improve causality extraction. The evaluation on a publicly available dataset demonstrated the advantage of our method on causality extraction. Causality extraction is a fundamental task. Therefore, besides the improvement of causality extraction, there are plenty of meaningful tasks for us to do, such as generating new causal hypothesis, generating scenario and so on. Meanwhile, causality extraction and its application research are full of challenges, including the credibility of causality in text, counterfactual causality, which need to be paid more attention.

Acknowledgments

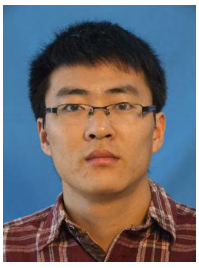
We are grateful to Prof. Wangxiang Che, Jiang Guo, Ruiji Fu and the anonymous reviewers for their insightful comments and suggestions. This work was supported by the National Key Basic Research Program of China via grant 2014CB340503 and the National Natural Science Foundation of China (NSFC) via grants 61133012 and 61472107.

References

[1] K. Radinsky, S. Davidovich, S. Markovitch, Learning causality for news events prediction, in: WWW '12, 2012, pp. 909–918.
[2] M. Riaz, R. Girju, Another look at causality: discovering scenario-specific contingency relationships with no supervision, in: 2010 IEEE Fourth International Conference on Semantic Computing (ICSC) 2010, pp. 361–368.
[3] C. Hashimoto, K. Torisawa, J. Kloetzer, M. Sano, I. Varga, J.-H. Oh, Y. Kidawara, Toward future scenario generation: extracting event causality exploiting semantic relation, context, and association features, in: ACL '14, 2014, pp. 987–997.
[4] R. Girju, Automatic detection of causal relations for question answering, in: Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering, 2003, pp. 76–83.
[5] A. Sorgente, G. Vettigli, F. Mele, Automatic extraction of cause-effect relations in natural language text, in: Proceedings of the 13th Conference of the Italian Association for Artificial Intelligence, 2013, pp. 37–48.
[6] A. Ittoo, G. Bouma, Minimally-supervised extraction of domain-specific part-whole relations using Wikipedia as knowledge-base, Data Knowl. Eng. 85 (2013) 57–79.
[7] M.A. Hearst, Automatic acquisition of hyponyms from large text corpora, in: Coling '92, 1992, pp. 539–545.
[8] F.M. Zanzotto, M. Pennacchiotti, M.T. Pazienza, Discovering asymmetric entailment relations between verbs using selectional preferences, in: ACL '06, 2006, pp. 849–856.
[9] J.-H. Oh, K. Torisawa, C. Hashimoto, M. Sano, S. De Saeger, K. Ohtake, Why-question answering using intra- and inter-sentential causal relations, in: ACL '13, 2013, pp. 1733–1743.
[10] S. Abe, K. Inui, Y. Matsumoto, Two-phased event relation acquisition: coupling the relation-oriented and argument-oriented approaches, in: Coling '08, 2008, pp. 1–8.
[11] Q.X. Do, Y.S. Chan, D. Roth, Minimally supervised event causality identification, in: EMNLP '11, 2011, pp. 294–303.
[12] C. Hashimoto, K. Torisawa, S. De Saeger, J.-H. Oh, J. Kazama, Excitatory or inhibitory: a new semantic orientation extracts contradiction and causality from the web, in: EMNLP '12, 2012, pp. 619–630.
[13] M. Riaz, R. Girju, In-depth exploitation of noun and verb semantics to identify causation in verb-noun pairs, in: 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2014, pp. 161–336.
[14] B. Rink, S. Harabagiu, Utd: classifying semantic relations by combining lexical and semantic resources, in: Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10, 2010, pp. 256–259.
[15] G. Zhou, J. Su, J. Zhang, M. Zhang, Exploring various knowledge in relation extraction, in: ACl '05, 2005, pp. 427–434.
[16] D. Zelenko, C. Aone, R. Richardella, Kernel methods for relation extraction, J. Mach. Learn. Res. 3 (2003) 1083–1106.
[17] R.J. Mooney, R.C. Bunescu, Subsequence kernels for relation extraction, in: NIPS, 2005, pp. 171–178.
[18] E. Pitler, A. Louis, A. Nenkova, Automatic sense prediction for implicit discourse relations in text, in: ACL '09, 2009, pp. 683–691.
[19] Z. Lin, M.-Y. Kan, H.T. Ng, Recognizing implicit discourse relations in the penn discourse treebank, in: EMNLP '09, 2009, pp. 343–351.
[20] Y. Hong, X. Zhou, T. Che, J. Yao, Q. Zhu, G. Zhou, Cross-argument inference for implicit discourse relation recognition, in: CIKM '12, 2012, pp. 295–304.
[21] X. Fang, G. Kennedy, Expressing causation in written English, RELC J. 23 (1) (1992) 62–80.
[22] R. Ji, L.-Y. Duan, J. Chen, L. Xie, H. Yao, W. Gao, Learning to distribute vocabulary indexing for scalable visual search, IEEE Trans. Multimed. 15 (1) (2013) 153–166.
[23] R. Ji, Y. Gao, R. Hong, Q. Liu, D. Tao, X. Li, Spectral-spatial constraint hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 52 (3) (2014) 1811–1824.
[24] E. Pitler, A. Nenkova, Using syntax to disambiguate explicit discourse connectives in text, in: ACL-IJCNLP '09, 2009, pp. 13–16.
[25] M. Collins, N. Duffy, Convolution kernels for natural language, in: NIPS, 2001, pp. 625–632.
[26] R. Lu, L. Xiang, M.-R. Liu, Q. Yang, Discovering topics from microblogs based on hidden topics analysis and text clustering, Pattern Recognit. Artif. Intell. 3, 2012.
[27] L. Jiang, H. Zhang, Z. Cai, A novel Bayes model: Hidden Naive Bayes, IEEE Trans. Knowl. Data Eng. 21 (10) (2009) 1361–1371.
[28] R. Ji, H. Yao, W. Liu, X. Sun, Q. Tian, Task-dependent visual-codebook compression, IEEE Trans. Image Process. 21 (4) (2012) 2282–2293.
[29] R. Ji, L.-Y. Duan, J. Chen, T. Huang, W. Gao, Mining compact bag-of-patterns for low bit rate mobile visual search, IEEE Trans. Image Process. 23 (7) (2014) 3099–3113.
[30] R. Ji, Y. Gao, W. Liu, X. Xie, Q. Tian, X. Li, When location meets social multimedia: a survey on vision-based recognition and mining for geo-social multimedia analytics, ACM Trans. Intell. Syst. Technol. (TIST) 6 (1) (2015) 1–18.
[31] R. Ji, L.-Y. Duan, J. Chen, H. Yao, J. Yuan, Y. Rui, W. Gao, Location discriminative vocabulary coding for mobile landmark search, Int. J. Comput. Vis. 96 (3) (2012) 290–314.
[32] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Mach. Learn. 29 (2-3) (1997) 131–163.
[33] B. Bohnet, J. Nivre, A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing, in: EMNLP '12, 2012, pp. 1455–1465.
[34] S. Abney, Part-of-speech tagging and partial parsing, in: Corpus-Based Methods in Language and Speech Processing, Kluwer Academic Publishers, Dordrecht, 1997, pp. 118-136.

Sendong Zhao is currently a Ph.D. candidate at Harbin Institute of Technology, Harbin, China. His research interests include natural language processing, knowledge reasoning, and text mining.

Ting Liu received his Ph.D. degree in 1998 from the Department of Computer Science, Harbin Institute of Technology, Harbin, China. He is a Full Professor in the Department of Computer Science, and the Director of the Research Center for Social Computing and Information Retrieval (HIT-SCIR) from Harbin Institute of Technology. His research interests include information retrieval, natural language processing, and social media analysis.



Sicheng Zhao is currently a Ph.D. candidate at Harbin Institute of Technology, Harbin, China. His research interests include affective computing, social media analysis and multimedia information retrieval.

Yiheng Chen received his Ph.D. degree in 2009 from the Department of Computer Science, Harbin Institute of Technology, Harbin, China. He is currently an assistant professor in the Department of Computer Science, Harbin Institute of Technology. His research interests include natural language processing, and social media analysis.

Jian-Yun Nie is a professor at the Université de Montréal, Canada. He has published more than 150 research papers in information retrieval and natural language processing in journals and conferences. He has served as a general cochair of the ACM-SIGIR conference in 2011. He is currently on the editorial board of seven international journals. He has been an invited professor and researcher at several universities and companies.