Transactions on Computational and Scientific Methods | Vo. 4, No. 10, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Causal Relationship Extraction Using BiGRU-Attention-CRF with Character Features and Iterated Dilated CNNs

Oberon Channing
Department of Computer Science, University of Delaware, USA
ococplmqaz@gmail.com

Abstract: Causality extraction is an essential task in natural language processing, aiming to identify and structure causal relationships from unstructured text. While deep learning techniques have advanced this field, challenges remain in extracting causal relationships effectively, particularly in representing lexical features and capturing multiple and overlapping causal relations. This paper proposes a novel method for causality extraction using a BiGRU-Attention-CRF model with enhanced feature representation. To improve the extraction performance, character features are integrated with pre-trained embeddings, and Iterated Dilated Convolutional Neural Networks (IDCNNs) are employed to capture local information while minimizing information loss. The BiGRU-Attention mechanism is utilized to extract deeper contextual information and improve training speed compared to traditional BiLSTM models. Experimental results demonstrate that the proposed method achieves competitive performance in causality extraction tasks, offering efficient training and robust feature extraction. However, the model's performance is limited by the dataset size, and it is currently confined to intra-sentence causality extraction. Future work will explore enhancing cross-sentence and document-level causality extraction using graph convolutional networks and low-resource question templates.

Keywords: Causality Extraction; Dilated Convolutions Networks; Gated Recurrent Unit; SequenceLabeling.

1. Introduction

Relation extraction is an essential research direction in natural language processing, which aims at extracting useful information from unstructured text and transforming it into structured information for use. Among the many relations of events, causality is one of the important semantic relations, as shown in Fig. 1, which shows the correspondence between events from before to after, from cause to effect, and due to the existence of a large amount of causal knowledge in natural language text, causality extraction has become an issue of vital concern in the field of artificial intelligence and has become more and more relevant in such tasks as information retrieval[1], event prediction, intelligent question and answer, text mining, and many other natural language processing tasks.

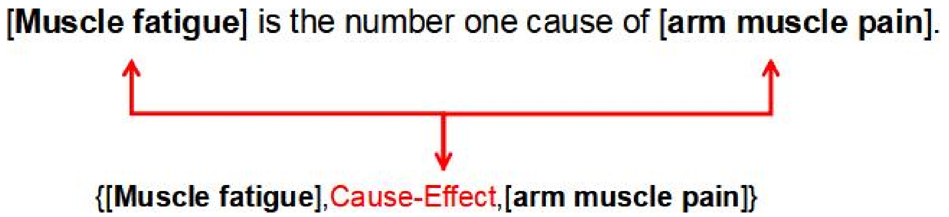
[Muscle fatigue] is the number one cause of [arm muscle pain].
{[Muscle fatigue],Cause-Effect,[arm muscle pain]}

Fig 1. A sentence expressing a causal relation, in this case, "muscle fatigue" is the cause and "arm

muscle pain" is the result of muscle fatigue

Causality is one of the more significant associations between events. Extracting causal relationships from texts has become a hot topic in natural language processing research. However, this research direction has not yet formed a mature research system, evaluation rules, and datasets for public evaluation. The lack of a unified causal sequence annotation method is one of the factors that hinder the progress of causal extraction research [2]. Most current studies focus on the causal relationship of a pair in an instance, but in reality, causality always exists in the form of one-to-one or one-to-many. In addition, causal extraction models often need to be trained with more than 10 million samples to reach the same level as humans. However, the known dataset size is much smaller than the desired value, and some of the datasets do not apply to all current causality extraction methods [3]. And most of the existing event causality extraction methods transform the extraction problem into a classification problem, where information features are extracted and then classified. Although this method was successful, the problem seems intractable in natural language processing tasks due to the ambiguity and diversity of natural language texts, the lack of contextualization of word features in causality extraction, and the inadequate representation of semantic features.

Traditional causality extraction methods in the past were divided into those based on pattern matching and those based on a combination of pattern matching and machine learning. The principle of pattern matching is to summarize the rules from massive linguistic texts and to construct a constraint template for causality extraction by using semantic features, syntactic features, and specific causal connectives such as cause, result, and lead. Although this method is intuitive to express, it needs to construct different forms of rule templates for various domains, which is less general and cannot balance the accuracy and recall rate. The approach combined with pattern matching and machine learning mainly divides the cause-effect extraction task into two sub-tasks: cause-effect pair extraction and relation classification, in which candidate cause-effect pairs are first identified through templates and then filtered according to syntactic or semantic features. It is more flexible than the former one but still requires a lot of labor and time cost in feature selection, and has high requirements on the size and quality of the corpus.

Due to the increasing forward development of the field of deep learning, attempts have begun to rely on the powerful representational capabilities of deep neural networks for causal relationship extraction. [4] determines whether there is a causal relationship between entities by using a convolutional neural network (CNN) to classify the relationship of a given entity. [5] proposed a knowledge-oriented CNN to classify events in texts causally by combining prior knowledge from a lexical knowledge base. [6] considered that the approach of using simple word embeddings to represent causal events ignores the inter-event context and the internal elements of the events, so the representation of events is enriched by multi-column convolutional networks to improve the causal event recognition performance. [7] proposed a new latent structure induction network to introduce an external knowledge base into the causal event recognition task and alleviate the problem of insufficient labeled data. Relationship classification methods often need to identify entities before determining the causal relationships between entities, and this method may generate redundant and redundant information, which affects the extraction efficiency. So, a new labeling scheme is proposed in the literature [8], where they convert causal relationship extraction into a sequential labeling task by using different end-to-end models on LSTM to label entities and inter-entity relationships concurrently. [9] considered the existence of interdependent associations of individual words in a sentence, so they enhanced the performance of causal extraction by constructing syntactic dependency graphs and assigning different weights to each word using the graph attention network (GAT) [10]. Although [8] used sequence annotation to accomplish the relationship extraction task for the first time, their labeling scheme cannot address the overlapping relationships that entities may contain among themselves in relationship extraction. Therefore, [11] designed a new tagging scheme and devised a tag2triplet algorithm to address the case of a sentence containing multiple causal triads and overlapping causal relations with the BiLSTM-CRF backbone using Flair embedding and the multi-head self-attention mechanism.

From the current survey, deep neural networks already achieve better results in solving the problem of causality extraction. This method can keep in long-distance information cases better dig deeper into text information. However, there are also some problems to be solved.

Firstly, the current study approach mostly simplified it as a relationship extraction problem, which judged whether there is a causality relationship between the entities according to the given candidate pairs in the clause, which merely classifies the relationship and cannot determine the causal direction between their entities.

Secondly, most works limited the study of causality to a single causal effect and less exploration of multiple causal effects in sentences. They rely on causal connectives to extract only the explicit causal relations with marks. In addition, the scope of causality extraction is limited to intra-sentence causality, unable to explore cross-sentence and cross-paragraph causal relationships.

Finally, there are several problems with the current approach, such as insufficient information about semantic features, word characteristics that do not match the contextual background, and inadequate representation of semantic features.

To alleviate the impact of the above problems, we use sequence labeling to extract causal relationships. And we enhance the characteristic representation by adding character features, together with the proposed BiGRU-Attention-CRF causal relationship extractor. We use Iterated Dilated Convolutional Neural Networks (IDCNNs) [13] to extract character characteristics with transfer embeddings and pre-trained word embeddings to enhance the feature representation. Subsequently, we input it into the model consisting of BiGRU and attention mechanism [20] to mine the text for deeper contextual information to capture the features in the sentence that are more important for causality. Then we input the feature information for label classification to the conditional random field (CRF) [21] by calculating the probability of adjacent labels to select the optimal sequence of labels.

The principal contributions of this paper are as follows.

(1) We add character characteristics to enrich the feature representation of the input text at different granularities and use IDCNNs to extract character characteristics to ensure retaining more features without losing information.

(2) We propose a neural network-based relational extractor BiGRU-Attention-CRF by using gated recurrent unit (GRU) instead of the common LSTM in sequence labeling to reduce model parameter count and improve training speed and model extraction performance in small-scale datasets.

2. Method

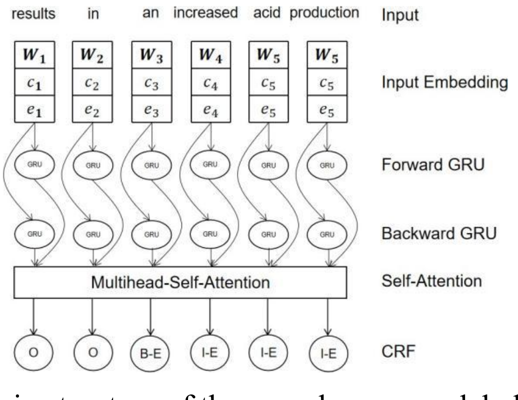

Fig 2. The main structure of the causal sequence labeling model

Fig. 2 illustrates the main structure of the causal draw model for the paper. We will present the parts of the models from top to bottom, taking the input sentenceS = $\{x_t\}_{t=}^{n}$ and the corresponding output tag sequence $y = \{y_i\}_{t=}^{n}$ as examples, where n is the length of the S.

2.1 IDCNNs

Character-level embeddings have been shown in [26][27] to improve the performance of such tasks relatively well. The advantage of using character-level features is that they can be extracted directly from the source text without designing additional manual features or preprocessing the original corpus.

In previous studies, CNN is commonly used to extract character features, which has some problems. Traditional convolutional networks obtain only a little information from the inputs after convolutional operations. So, for capturing more contextual information, it is necessary to add more convolutional layers, which increases network layers and parameters and leads to the overfitting risk, and repeated pooling operations to integrate information will cause some information loss. Therefore, dilated convolutions (DCNN) are proposed [12]. In conventional convolutional neural networks, the convolution kernel slides in the continuous region, while the expansion convolution adds an expansion width on this basis. During the convolution operation, the data in the middle of the expansion width will be skipped, and a broader input matrix data will be obtained under the condition that the size of the convolution kernel remains unchanged, thus increasing the perception field of the convolution kernel.

Although more features are obtained by superimposing the number of expanded convolutional layers, the corresponding increase in the number of parameters leads to problems such as convergence difficulties and over-fitting during the actual training. Therefore, [13] proposes Iterated Dilated Convolutional Neural Networks (IDCNNs). By using the same dilated convolution (Fig. 3) many times, each iteration takes the last result as input, and reusing the same parameters circularly provides a wide effective input width and desirable generalization capability.

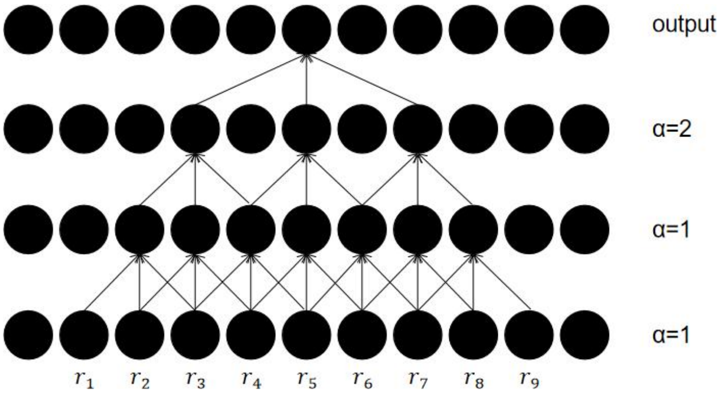

Fig 3. DCNN architecture for recycling in IDCNNS.

We use IDCNNs to extract character features by superimposing four inflated convolutional blocks of the same size. the structure of the convolutional blocks is shown in Fig.3, and the expansion coefficients of DCNN within each convolutional block are α as 1,1,2. The network takes the corresponding vector of each character as input and outputs the character features corresponding to each word after iterations of the inflated convolution.

2.2 Contextual String Embeddings

The word embedding representations can be obtained by training neural network language models, but the obtained embeddings are fixed embeddings that cannot characterize the multiple meanings of words, and it would be difficult to obtain rich word embedding representations if there is insufficient corpus data. As research advances, [14] and [15] address the performance limitations of neural network models under corpus insufficiency by training on large unlabeled corpora to obtain pre-trained language models containing text-rich semantic information representations.

[16] used recurrent neural networks for language modeling, modeling words, and contexts as sequences of characters, which are used to handle some rare words and misspelled words and word structures as prefixes and endings. And the word meanings are captured according to the context so that the words are entangled in different embeddings in different contexts and finally pre-trained in a large unlabeled corpus to refer to the obtained word representations as contextual string embeddings.

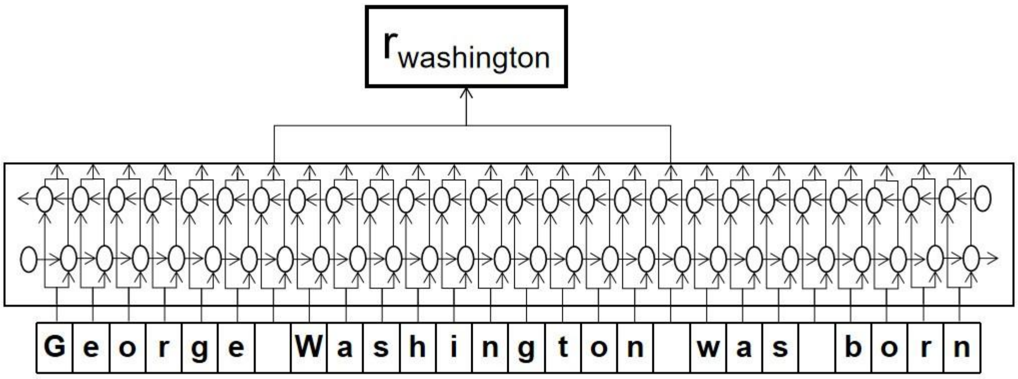

Fig 4. Extract context string embedding of the word in the context of the sentence (" Washington ").

The language model is shown in Fig. 4, [16] chooses a variant of the recurrent neural network, LSTM, as the basic architecture of the model, which consists of a forward language model and a negative reverse language model. An individual character in a word is used as the basic unit of the model input by using it as the basic unit, so each unit in the character sequence can be used to be trained to predict the next word. Taking the word "Washington" as an example, the forward LSTM language model extracts the character features of each word from left to right from the beginning of the sentence and passes them backward, and then extracts the output hidden state h^{f}_{end+} after the last character of the word, which contains information from the beginning of the sentence to the word. Similarly, the reverse LSTM language model passes from back to front and outputs the hidden state h^{b}_{start-} before extracting the first character of the word, which contains the information from the end of the clause to the word. The two output hidden states are then concatenated to get the final context string embedded w^{CharLM}_{i}.

$$w^{CharLM}_{i} = [h^{f}_{end+1}, h^{f}_{end+1}]$$

Finally, we input the contextual string embeddings, the character features, and the trained word embeddings from [17] for stitching into the BiGRU network for further feature extraction.

2.3 BiGRU

To solve the problem that RNN is prone to gradient disappearance and gradient explosion when dealing with long-distance dependency, [18] and [19] proposed long-short Term Memory (LSTM) and Gated recurrent unit (GRU). Compared with the gating structure of the input gate, forgetting gate, and output gate of LSTM, the gated loop unit GRU (Fig. 5) loses the cell state, directly uses the hidden state H_t to transmit information, fuses the input gate and forgetting gate into update gate z_t, and changes the output gate into reset gate r_t. Compared with LSTM networks, GRU structure is relatively easy, less prone to overfitting, easier to converge with smaller data sets and requires fewer iterations, and has the same functions as LSTM network models. Therefore, GRU has fewer parameters and a shorter training time than LSTM with the same performance.

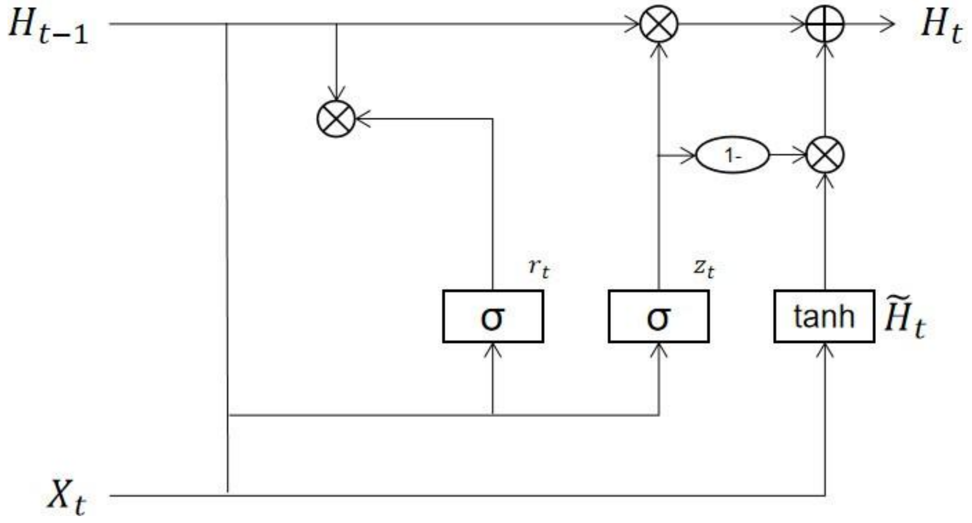

Fig 5. Internal structure of gated loop unit GRU

Generally, the process of input $X_t = [r_i, e_i, w_i^{CharLM}]$ into the GRU unit to obtain a hidden state at time T is as follows:

$$z_t = \sigma(W_z X_t + U_z H_{t-1} + b_z)$$

$$r_t = \sigma(W_r X_t + U_r H_{t-1} + b_r)$$

$$\widetilde{H}_t = \tanh(W_h X_t + U_h(H_{t-1} \otimes r_t) + b_n)$$

Where W_z、 W_r、 W_h and U_z 、 U_r 、 U_h are the weight matrix of input and hidden state at the current time t; b_z、 b_r、 b_n is the bias weight; σ represents the sigmoid function; x is Hadamardproduct.

2.4 Multi-head Self-attention

To obtain the long-distance dependence in sentences, we connect multi-directional self-attention [20] to the tail of BiGRU, and weighted the contextual semantic features obtained by BiGRU. The Attention mechanism is similar to that when humans observe objects, they selectively pay Attention to some information while ignoring other information. Therefore, the Attention mechanism captures more important features by weighting input vectors.

Specifically, H is the output of BiGRU, and the essence of Multi-head Self-attention is to project the vector H n times to generate different Q_i, K_i, V_i query, key, and value matrices. Then the different matrices are mapped to different subspaces for attention computation separately, and the computed results $head_i$ of different subspaces are then merged and operated to obtain the final output M by a single linear transformation.

$$Q_i = HW_i^Q$$

$$K_i = HW_i^K$$

$$V_i = HW$$

Where W is the transformation matrix,$W_{i,j}$ denotes the label transfer score, and y_O and y_n represent the unique labels at the beginning and end of the sentence. Then the probability of label sequence y in the case of a given sentence S is expressed as:

$$p(y|S) = \frac{e^{score(S,y)}}{\sum_{y \in Y_S} e^{score(S,y)}}$$

The maximum likelihood function for the training process is as follows:

$$\log(p(y|S)) = score(S, y) — \log(\sum_{y \in Y_S} e^{score(S,y)})$$

Where Y_s denotes all potential tagged sequences of sentence S. Finally, the predicted sequence label with the highest conditional probability to obtain:

$$y^* = argmaxscore(S, y)$$

3. Experiments

3.1 Dataset

In this experiment, the extended SemEval 2010 Task 8 dataset was used, which contained 5,236 sentences in total, among which 1,270 sentences contained causality. The training set consists of 4450 sentences containing 157 causal triples in total. The test set consists of 804 sentences containing 296 causal triples.

Table 1. Statistics of different types of tags.

	B-C	I-C	B-E	I-E	B-Emb	I-Emb	O
Train	1308	1421	1268	1230	55	55	66614
Test	236	229	238	230	9	16	12784

3.2 Hyperparameters

The model was built by TensorFlow 2.3, using a word vector of dimension 300 trained in the literature[17]. The character features are obtained by uniformly distributed random initialization, using pre-trained contextual character embeddings from the literature [16]. For IDCNNs layer iteration block DCNN number of layers set to 3, its iteration number to 4, expansion coefficients α to 1, 1, 2. its hidden state size set to 256 in GRU, dropout set to 0.5. the number of heads set to 3 in multi-headed self-attention, size to 8. Batch size set to 64, iteration number set to 180, learning rate to 0.002, and use Nadam as the optimizer.

3.3 Baselines

For an adequate comparison of the effectiveness of the methods in this paper, we use several classical causality extraction methods to compare with our approach. These methods can fall into two categories: pipeline methods and sequence labeling methods.

Rules-Bayesian [22]:Pattern matching takes place according to the set rule template, firstly extracting candidate causal pairs from the text, and then filtering non-causal pairs using a Bayesian classifier and Laplace smoothing.

CausalNet [23]:The method identifies causal relationships between two arbitrary short texts by causal strength (CS). To facilitate comparison, we then add the same causality extraction module as in [22] to [23]. The comparison is performed by calculating the CS scores of the candidate causal triads and a set threshold value φ. If CS(c,e) > φ, a causal relationship is considered to exist, and vice versa.

The baseline for BiLSTM based model is as follows:

BiLSTM-softmax[24]:The model has two parts, the BiLSTM encoder and the softmax classifier.

BiLSTM-CRF[25]:A classical choice of sequence annotation task, consisting of a BiLSTM encoder and a CRF classifier.

CNN-BiLSTM-CRF[26]:A hierarchical BiLSTM-CRF model with character-level features extracted by a character-level CNN encoder and character embeddings connected to their pre-trained word embeddings input to the BiLSTM.

CLSTM-BiLSTM-CRF[27]:A character level embedding using a character LSTM encoder (CLSTM) instead of CNN to learn character level embedding similar to the hierarchical BiLSTM-CRF model.

BERT-CISAN[28]:Considering the priors knowledge in different domains affects causality extraction, the model uses BERT as an encoding layer of BiLSTM-CRF and uses the convolutional network with pre-determined weights to extract features. It then uses a key query attention mechanism to reduce incorrect causal candidate pairs.

The following model adds additional contextual string embedding with Multi-head Self-attention and is baselined with the BiLSTM-CRF architecture:

BiLSTM-Attention-CRF:This model adds self-attention to the base to enhance long-range dependencies and enriches the feature representation by adding contextual string embeddings.

CNN-BiLSTM-Attention-CRF [11]: This model adds a hierarchical structure to BiLSTM-Attention-CRF to extract character features by character-level CNN.

PosNet[29]:Pointer labeling first constructs text features containing location information, then creates two start and end Pointers to predict the start and end positions of causal entities in the sentence. The obtained causal entity positions to extract the causal entities by the assembly algorithm.

To validate the performance of our causality extraction using IDCNNs to extract character features and BiGRU-Attention-CRF, we chose to extract character features using a single CNN as a comparison to evaluate the effectiveness of our proposed method when other experimental parameters are consistent.

BiGRU-Attention-CRF: As with the BiLSTM-Attention-CRF structure, a simpler internal structure and faster GRU are chosen to extract contextual features.

CNN-BiGRU-Attention-CRF: The character features used CNN to extract character features based on BiGRU-Attention-CRF.

3.4 Experimental Results

The performance of different models on causality extraction shows in Table 2. IDCNNs-BiGRU-Attention-CRF outperforms all other models in the test set with an F_1 value of 81.06%. this proves the effectiveness of our proposed method. It also further illustrates that for causality extraction, the sequence labeling method outperforms the pipeline method and slightly outperforms the pointer labeling method. And with more causal pairs in the sentence, our approach will recognize faster than [29].

Table 2. Comparison of scores with current causality extraction methods

Models	F1
CausalNet	57.61%
Rules-Bayesian	59.59%
BiLSTM-softmax	73.33%
CNN-BiLSTM-CRF	75.57%
CLSTM-BiLSTM-CRF	75.06%
BiLSTM-CRF	76.78%
BERT-CISAN	77.65%
BiLSTM-Attention-CRF	78.42%
CNN-BiLSTM-Attention-CRF	79.44%
BiGRU-Attention-CRF	79.20%
CNN-BiGRU-Attention-CRF	79.31%

PosNet	80.90%
IDCNNs-BiGRU-Attention-CRF	**81.06%**

In addition, as shown in Table 2, the models' performance is significantly improved by 1.84% and 2.66% after adding contextual string embedding and self-attention to the BiLSTM-CRF architecture. That indicates that the contextualized character-level word embedding is more suitable for the causality extraction task, and the addition of Multi-head Self-attention can effectively extract more causal features.

Table 3. Performance scores of the model with two different methods of extracting character features

Models	P	R	F1
BiLSTM-Attention-CRF	76.47%	80.68%	78.42%
CNN-BiLSTM-Attention-CRF	77.98%	**80.95%**	79.44%
BiGRU-Attention-CRF	**83.60%**	75.42%	79.20%
CNN-BiGRU-Attention-CRF	82.14%	76.67%	79.31%
IDCNNs-BiGRU-Attention-CRF	81.64%	80.48%	**81.06%**

For further verification of the effectiveness of IDCCNs-BiGRU-Attention-CRF, we selected other sequence labeling models to compare in detail. Also, it is seen from Table 3 that without adding character features, the BiGRU-Attention-CRF score is 79.2%, while the BiLSTM-Attention-CRF score is 78.42%. With the addition of a single CNN to extract character features, it is 79.31% and 79.44%, respectively. Thus, it is demonstrated that adding character features has an impact on the performance of model extraction under a specific task. After extracting character features using IDCNNs, the recall R of BiGRU-Attention-CRF improved by 5.06% and 3. 86%, and although the accuracy of P decreased, F improved by 1.84% compared with no character features added and by 1.73% compared with character features extracted using CNNs, and F_1 reached 81.06%. These results indicate that compared with the character features extracted by CNN, IDCNNs can learn more local upper and lower character features by expanding convolution without losing information, which can make the character feature representation more complete, enrich the feature representation of words with different granularity, and help improve the performance of causality extraction.

Table 4. Different models with the number of parameters and training time

Models	Total Parameters	Average time for an epoch
BiLSTM-Attention-CRF	14,232,306	18s362ms
CNN-BiLSTM-Attention-CRF	14,298,546	21s419ms
BiGRU-Attention-CRF	11,851,506	16s322ms
CNN-BiGRU-Attention-CRF	11,902,386	18s362ms
IDCNNs-BiGRU-Attention-CRF	11,935,146	18s372ms

Adding character features improves the scores of models, yet it also brings about the problem of increasing the parametric. As Table 4 shows, compared to no additional character features, the model reduces around 0.8M parameters and saves 3s in training an epoch, which indicates that adding character features makes the model structure complex and takes more time for training, but the model performance improves accordingly. Secondly, compared with different networks to extract character characteristics, we find that using IDCNNs increases parametric by about 0.5M while the training time remains the same. That indicates the benefit of our method while increasing parametric is worthwhile. We compare the experimental results of BiLSTM and BiGRU models and find that both models with GRU perform slightly better than LSTM, probably since GRU has fewer parameters than LSTM caused by the structural difference, which makes it less dependent on the training set size and easier to converge. Therefore, without affecting performance, BiGRU can save more time in training.

3.5 Analysis and Discussion

Table 5 shows the performance of our proposed model for identifying different labels on the SemEval 2010 Task 8 dataset. The model performs well in identifying the cause (C) and effect (E). For the cause label (C), the P value was 89.28%, the R was 66.24%, and the F_1 was 76.05%. For the result label (E), the P value was 90.94%, the R was 62.18%, and the F_1 value was 73.68%. However, for the embedded causality (Emb) label, this model is not ideal, and the recall rate R is 8% and F_1 is only 14.81%.

Table 5. Performance scoring results of different models for different labels

Models	Type	P	R	F1
IDCNNs-BiGRU-Attention-CRF	C	0.8928	0.6624	0.7605
	E	0.9094	0.6218	0.7368
	Emb	1.000	0.0800	0.1481

4. Conclusion

In response to the problem of incomplete representation of lexical features in existing causality extraction studies, we propose a causality extraction method of BiGRU-Attention-CRF with fused character features. Different from the previous extraction of character features by single CNN, we obtain more local information by IDCNNs with reduced information loss and combine the pre-trained contextual string embedding and word embedding into BiGRU-Attention-CRF for sequence annotation to complete causality extraction. The experimental results show that IDCNNs are more favorable for extracting character features than CNN and using BiGRU-Attention-CRF achieve a shorter training time than BiLSTM-Attention-CRF for causal relationship extraction without affecting F_1.However, owing to the small dataset size, the model's performance is limited, and the current methodis limited to extracting intra-sentence causality, unable to achieve cross-sentence or cross-segment causality extraction. Therefore, in future work, we will consider using the language model to improvethe performance of causality extraction by constructing question templates with low resources. Secondly, for document-level causality extraction, we try to apply graph convolutional networks to extract causality.

References

[1] She X, Chen J, Chen G. Joint Learning With BERT-GCN and Multi-Attention for Event Text Classification and Event Assignment[J]. IEEE Access, 2022, 10: 27031-27040.

[2] Yang J, Han S C, Poon J. A survey on extraction of causal relations from natural language text[J]. Knowledge and Information Systems, 2022: 1-26.

[3] Xu J, Zuo W, Liang S, et al. A review of dataset and labeling methods for causality extraction [C]// Proceedings of the 28th International Conference on Computational Linguistics. 2020: 1519-1531.

[4] De Silva T N, Zhibo X, Rui Z, et al. Causal relation identification using convolutional neural networks and knowledge-based features[J]. International Journal of Computer and Systems Engineering, 2017, 11 (6): 696-701.

[5] Li P, Mao K. Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts[J]. Expert Systems with Applications, 2019, 115: 512-523.

[6] Jin G, Zhou J, Qu W, et al. Exploiting Rich Event Representation to Improve Event Causality Recognition [J]. INTELLIGENT AUTOMATION AND SOFT COMPUTING, 2021, 30(1): 161-173.

[7] Cao P, Zuo X, Chen Y, et al. Knowledge-enriched event causality identification via latent structure

induction networks[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 4862-4872.

[8] Zheng S, Wang F, Bao H, et al. Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme [C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 1227-1236.

[9] Jinghang X, Wanli Z, Shining L, et al. Causal relation extraction based on graph attention networks[J]. J. Comput. Res. Dev, 2020, 57: 159-174.

[10] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[J]. arXiv preprint arXiv: 1710. 10903, 2017.

[11] Li Z, Li Q, Zou X, et al. Causality extraction based on self-attentive BiLSTM-CRF with transferred embeddings[J]. Neurocomputing, 2021, 423: 207-219.

[12] Yu F, Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions[C]//ICLR. 2016.

[13] Strubell E, Verga P, Belanger D, et al. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 2670-2680.

[14] PETERS M E, NEUMANN M, IYYER M, et al. Deep Contextualized Word Representations, New Orleans, Louisiana, F, 2018 [C]. Association for Computational Linguistics.

[15] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[16] Akbik A, Blythe D, Vollgraf R. Contextual string embeddings for sequence labeling[C]//Proceedings of the 27th international conference on computational linguistics. 2018: 1638-1649.

[17] Komninos A, Manandhar S. Dependency based embeddings for sentence classification tasks[C]// Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. 2016: 1490-1500.

[18] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[19] Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [J]. arXiv preprint arXiv:1412.3555, 2014.

[20] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[21] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J]. proceedings of icml, 2002.

[22] Sorgente A, Vettigli G, Mele F. Automatic extraction of cause-effect relations in Natural Language Text[J]. DART @ AI* IA, 2013, 2013: 37-48.

[23] Luo Z, Sha Y, Zhu K Q, et al. Commonsense causal reasoning between short texts[C]//Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning. 2016.

[24] Wang P, Qian Y, Soong F K, et al. Part-of-speech tagging with bidirectional long short-term memory recurrent neural network[J]. arXiv preprint arXiv:1510.06168, 2015.

[25] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

[26] Ma X, Hovy E. End-to-end sequence labeling via Bidirectional lstm-cnns-crf[J]. arXiv preprint arXiv: 1603. 01354, 2016.

[27] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[J]. arXiv preprint arXiv:1603.01360, 2016.

[28] Wang Z, Wang H, Luo X, et al. Back to Prior Knowledge: Joint Event Causality Extraction via Convolutional Semantic Infusion[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Cham, 2021: 346-357.

[29] Guangli Z, Xin X, Shunxiang Z, Houyue W, Ju H, et al. PosNet: Position-based Causal Relation Extraction Network [J/OL]. Computer Science:1-11[2022-08-28]. http:// kns.cnki.net/ kcms/detail/ 50.1075. TP. 2022 0810. 0911.014.html

[30] Cui L, Wu Y, Liu J, et al. Template-based named entity recognition using BART[J]. arXiv preprint arXiv: 2106. 01760, 2021.

[31] Phu M T, Nguyen T H. Graph convolutional networks for event causality identification with rich

document-level structures[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 3480-3490.