

# Causality Extraction Based on Self-Attentive BiLSTM-CRF with Transferred Embeddings

Zhaoning Li, Qi Li, Xiaotian Zou, Jiangtao Ren\*

*School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, Guangdong,  
P.R.China 510006*

---

## Abstract

Causality extraction from natural language texts is a challenging open problem in artificial intelligence. Existing methods utilize patterns, constraints, and machine learning techniques to extract causality, heavily depending on domain knowledge and requiring considerable human effort and time for feature engineering. In this paper, we formulate causality extraction as a sequence labeling problem based on a novel causality tagging scheme. On this basis, we propose a neural causality extractor with the BiLSTM-CRF model as the backbone, named **SCITE** (**S**elf-attentive BiLSTM-**CRF** w**I**th **T**ransferred **E**mbellings), which can directly extract cause and effect without extracting candidate causal pairs and identifying their relations separately. To address the problem of data insufficiency, we transfer contextual string embeddings, also known as Flair embeddings, which are trained on a large corpus in our task. In addition, to improve the performance of causality extraction, we introduce a multihead self-attention mechanism into SCITE to learn the dependencies between causal words. We evaluate our method on a public dataset, and experimental results demonstrate that our method achieves significant and consistent improvement compared to baselines.

**Keywords:** Causality extraction, Sequence labeling, BiLSTM-CRF, Flair

---

\*Corresponding author

*Email addresses:* lizhn7@mail2.sysu.edu.cn (Zhaoning Li),  
liqi38@mail2.sysu.edu.cn (Qi Li), zouxt5@mail2.sysu.edu.cn (Xiaotian Zou),  
issrjt@mail.sysu.edu.cn (Jiangtao Ren)

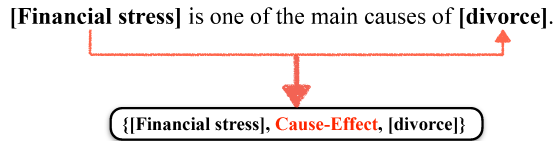


Figure 1: A sentence expressing causal relations, in this case, “financial stress” is the cause, “divorce” is the effect caused by financial stress.

embeddings, Self-attention

## 1. Introduction

Natural language text contains considerably causal knowledge, as shown in Fig. 1. In recent years, causality extraction has become increasingly important for many natural language processing tasks, such as information retrieval [1, 2], event prediction [3, 4], question answering [5, 6, 7], generating future scenarios [8, 9], decision processing [10], medical text mining [11, 12, 13] and behavior prediction [14]. However, due to the ambiguity and diversity of natural language texts, causality extraction remains a hard NLP problem to solve.

Traditional methods for causality extraction can be divided into two categories: methods based on patterns [1, 11, 15, 16] (Section 5.1), and methods based on a combination of patterns and machine learning techniques [5, 17, 18, 19] (Section 5.2). The former often has poor cross-domain applicability, fails to balance precision and recall and may require extensive domain knowledge to solve problems in a particular area. The latter usually requires considerable human effort and time on feature engineering, relying heavily on the manual selection of textual features. Generally, it divides causality extraction into two subtasks, candidate causal pairs extraction and relation classification (filtering noncausal pairs). The results of candidate causal pairs extraction may affect the performance of relation classification and generate cascading errors.

Zheng et al. [20] first proposed a tagging scheme that makes it possible for models to extract entities and relations simultaneously. In their tagging scheme, they apply a cartesian product of the entity mention tags and the relation type

tags, and then assign a unique tag that encodes entity mentions and relation types for each word. Inspired by their novel idea, we focus on a causal triplet that is composed of two event entities and their relation. For instance, the sentence in Fig. 1 contains a causal triplet: “{**financial stress**, **cause-effect**, **divorce**}”. Thus, we can model the causal triplets directly, rather than breaking causality extraction into two subtasks. Based on the motivations, we formulate causality extraction into a sequence tagging problem and propose a causality tagging scheme (Section 2.1) to achieve direct causality extraction. However, the tagging scheme proposed by Zheng et al. [20] cannot identify the overlapping relations in a sentence; it only considers situations where an entity belongs to one triplet: if an entity participates in multiple relations, its tag should not be unique. To address this problem, we design a **tag2triplet** algorithm (Section 2.2) to handle multiple causal triplets and embedded causal triplets in the same sentence. Finally, we combine the causality tagging scheme with a deep learning architecture (Section 2.3) to minimize feature engineering while efficiently modeling causal relations in natural language text.

We notice that some researchers have also proposed deep learning technique-based methods for causality extraction in recent years (Section 5.3). Although their works are commendable, some works [21, 22, 23, 24] are only a classification of causal relations rather than an extraction of complete causal triplets, and others [25, 26] mainly focus on the identification of the linguistic expressions for causality instead of the commonsense causality extraction in this paper.

By applying our causality tagging scheme, we use the model based on BiLSTM-CRF [27] to extract causal triplets directly. However, we find that two obstacles hinder the further improvement of the performance of the deep learning model.

**First**, it is difficult to train a superior deep learning model without any prior knowledge in the case of data insufficiency in the existing corpus [28, 29, 30]. To alleviate this problem, we incorporate Flair embeddings [31] into our task, which use the internal states of a character language model trained on a large corpus to create word embeddings (Section 2.3.2). Experimental results show that this contextual string embedding that has initiated a new technology trend

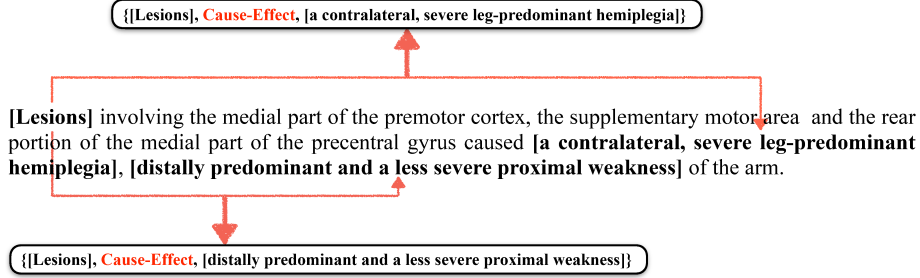


Figure 2: The second causal triplet: “[{lesions}, cause-effect, [distally predominant and a less severe proximal weakness]]” spans almost the entire sentence.

in NLP can drastically improve the performance of causality extraction.

**Second**, in terms of their positions in the text, cause and effect are sometimes far from each other, as Fig. 2 shows. The long-range dependency in the causal triplet creates difficulty and ambiguity in the deep learning model, but a set of logical rules based on dependency trees can easily and accurately extract such triplets. To learn this kind of long-range dependency between cause and effect, we introduce the multihead self-attention mechanism [32] into our model (Section 2.3.4). Unlike the LSTM-based model that recursively processes each word, the self-attention mechanism can conduct direct connections between two arbitrary words in a sentence and thus allows unimpeded information flow through the network [33].

The contributions of this paper can be summarized as follows:

1. We design a novel causality tagging scheme to directly extract causalities in texts and can easily transform the causality extraction into a sequence labeling task and handle multiple causal triplets and embedded causal triplets in the same sentence.
2. Based on our causality tagging scheme, we propose **SCITE** (**S**elf-attentive **B**iLSTM-**C**RF **w**ith **T**ransferred **E**MBEDDINGS), a neural-based causality extractor with transferred contextual string embeddings trained on a large corpus. To the best of our knowledge, we are the first to transfer Flair embeddings into causality extraction.

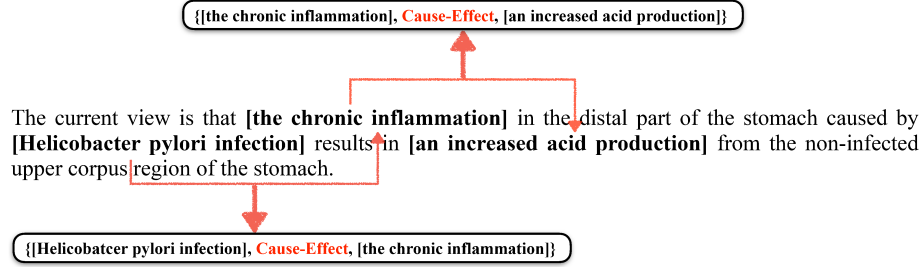


Figure 3: Two causal triplets share the same causal event entity: “the chronic inflammation” within a sentence.

3. We introduce the multihead self-attention mechanism into SCITE, which enables the model to capture long-range dependencies between cause and effect.
4. Extensive experimental results (Section 3) and further analysis (Section 4) show that our method achieves significant and consistent improvement compared to other baselines. We release the code and dataset to the research community for further research <sup>1</sup>.

## 2. Method

### 2.1. Causality Tagging Scheme

We use the “BIO” (begin, inside, other) and “C, E, Emb” (cause, effect, embedded causality) signs to represent the position information of the words and the semantic roles of the causal events, respectively, where **embedded causality** [30] indicates that a causal event has different roles of causality in different triplets. Fig. 3 is an example of an embedded causality in a sentence. The example sentence contains two causal triplets: “{**the chronic inflammation**, **cause-effect**, **an increased acid production**}” and “{**Helicobacter**, **cause-effect**, **the chronic inflammation**}”, note that “the chronic inflammation” is the cause in the first triplet and the effect in the second triplet.

<sup>1</sup><https://github.com/Das-Boot/scite>

Input Sentence: The current view is that the chronic inflammation in the distal part of the stomach caused by *Helicobacter pylori*  
Tag Sequence: O O O O O B-Emb I-Emb I-Emb O O O O O O O O O B-C I-C

Input Sentence: infection results in an increased acid production from the non-infected upper corpus region of the stomach.  
Tag Sequence: I-C O O B-E I-E I-E I-E O O O O O O O O

Figure 4: A standard annotation for the example sentence based on our causality tagging scheme.

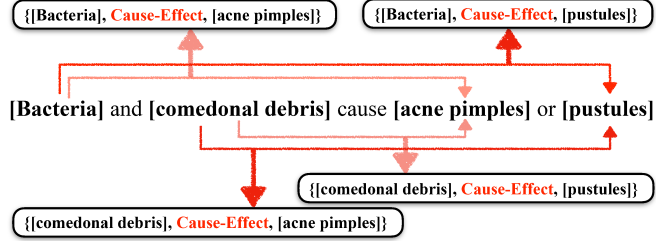


Figure 5: For any one of the four causal triplets in the above sentence, there is another causal triplet sharing the same cause or effect.

Fig. 4 shows an example of such causality sequence tagging. Based on our causality tagging scheme, we label the causal event entities “chronic inflammation”, “*Helicobacter pylori* infection” and “increased acid production” separately with our special tags. Specifically, tag “O” represents the “other”, which means that the corresponding word is irrelevant in any causality components. Tag “B-C” represents the “cause begin”, tag “I-C” represents the “cause inside”, tag “B-E” represents the “effect begin”, tag “I-E” represents the “effect inside”, tag “B-Emb” represents the “embedded causality begin”, and tag “I-Emb” represents the “embedded causality inside”. Thus, the total number of tags is  $N_t = 7$ .

## 2.2. From Tag Sequence to Causal Triplets

We design a **tag2triplet** algorithm for automatically obtaining the final extracted triplets from the tag sequence in Fig. 4. To better illustrate this algorithm, we define two types of causality: simple causality and complex causality.

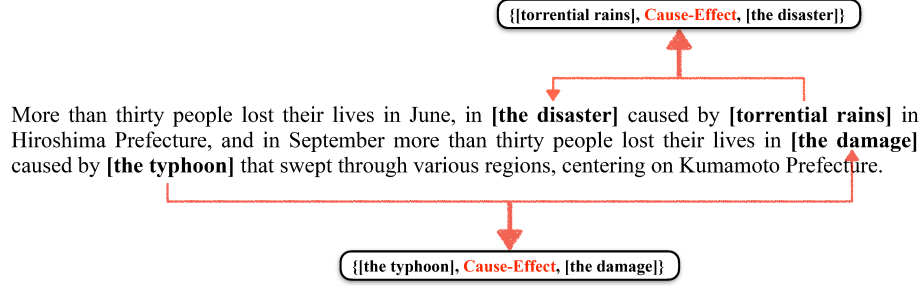


Figure 6: The causal triplet “{**torrential rains**, **cause-effect**, **the disaster**}” and “{**the typhoon**, **cause-effect**, **the damage**}” do not share the same cause or effect.

### 2.2.1. The Case of Simple Causality

Simple causality can be classified into two types:

1. There is only one cause or one effect in the sentence, and there is no embedded causality, that is,  $N_C = 1$  or  $N_E = 1$  and  $N_{Emb} = 0$ , where  $N_C$ ,  $N_E$ , and  $N_{Emb}$  respectively indicate the number of tags “B-C”, “B-E”, and “B-Emb” in the sentence. The example sentences in Fig. 1 and Fig. 2 are both of this type of causality.
2. There are multiple causes and effects in the sentence, and there is no embedded causality, i.e.,  $N_C > 1$ ,  $N_E > 1$  and  $N_{Emb} = 0$ . In addition, for each causal triplet in the sentence, there must be at least one causal triplet that shares the same cause or effect. The example sentence in Fig. 5 is this type of causality.

### 2.2.2. The Case of Complex Causality

Complex causality has the following two types:

1. There is embedded causality in the sentence, that is,  $N_C > 0$ ,  $N_E > 0$  and  $N_{Emb} > 0$ . The example sentence in Fig. 3 is this type of causality.
2. There are multiple causes and effects in the sentence, and there is no embedded causality, i.e.,  $N_C > 1$ ,  $N_E > 1$  and  $N_{Emb} = 0$ . In addition, in all the causal triplets in the sentence, there must be at least one causal triplet that does not share the same cause or effect with any other triplets.

The example sentence in Fig. 6 is this type of causality. Note that the distribution of causality in the sentence of Fig. 5 is different from that in Fig. 6: each causal triplet in the former is mixed together, and each causal triplet in the latter is separated.

### 2.2.3. Tag2triplet Algorithm

The tag2triplet algorithm is described in Algorithm 1. We elaborate on the tag2triplet algorithm by taking the sentence  $S$  and its corresponding tag sequence  $S_{tag}$  as an example in Fig. 4; the intermediate result is shown in Table 1.

First, we count the out-degree and in-degree for causality and find the index of causality in  $S_{tag}$ . Specifically, the out-degree of “cause” is recorded as 1, the in-degree of “effect” is recorded as 1, and the out-degree and in-degree of “embedded causality” are both recorded as 1. Then, we determine whether the  $S$  is simple causality or complex causality according to the number and the distribution of each causal tag: “C”, “E” and “Emb”. In  $S_{tag}$ ,  $N_{Emb} = 1$ , and thus,  $S$  is complex causality. Then, we apply a **Cartesian product** of the causal entities composed of causal tags to generate the candidates of the causal triplet. In Table 1, the candidate “ $(E_0, E_2)$ ” represents the causal triplet “{**the chronic inflammation, cause-effect, an increased acid production**}”.

Next, **combination** returns  $i$  length subsequences of triplets from the input candidates. After that, we determine whether the out-degree and in-degree of each combination of candidates are consistent with the original out-degree and in-degree of  $S_{tag}$ . Then, we determine whether the combination matches the rules according to the coordinating conjunction in  $S$ , for example, if there is a coordinating conjunction “and” between adjacent causes in the same clause, then the two causes will form their respective causal triplets with the same effect, as in the “bacteria” and “comedonal debris in Fig. 5. Finally, we select the combination with the shortest distance from the combinations that passed the checks as the extracted causal triplets. Since only one combination “ $(E_0, E_2)$ ,  $(E_1, E_0)$ ” passed all the checks, we directly output it as the final result.



```

input : A tag sequence  $S_{tag}$  corresponding to sentence  $S$ 
output: The causal triplets in sentence  $S$ 

1 Count the out-degree and in-degree for the causality in the  $S_{tag}$ ;
2 Find the index of causality  $idx$  in the  $S_{tag}$ ;
3 if Causality in  $S \in$  simple causality then
4   candidate  $\leftarrow$  CartesianProduct( $idx$ );
5   if CheckConjunction(candidate,  $idx$ ,  $S$ ) is true then
6     causal triplets  $\leftarrow$  candidate;
7   end
8 end
9 if Causality in  $S \in$  Complex Causality then
10  candidates  $\leftarrow$  CartesianProduct( $idx$ );
11  for  $i \leftarrow$  Max(Sum(out-degree), Sum(in-degree)) to Len(candidates)
12    do
13      flag  $\leftarrow$  0;
14      records  $\leftarrow$  [ ];
15      for  $j \in$  Combination(candidates,  $i$ ) do
16        if CheckDegree( $j$ , out-degree, in-degree) is true and
17          CheckConjunction( $j$ ,  $idx$ ,  $S$ ) is true then
18            distance  $\leftarrow$  SumDistance( $j$ ,  $idx$ );
19            AppendToRecord( $j$ , distance);
20            flag  $\leftarrow$  1;
21          end
22        end
23      if flag  $\neq$  0 then
24        break;
25      end
26    end
27  causal triplets  $\leftarrow$  Min(records, key=records[-1]);
28 end

```

**Algorithm 1:** Tag2triplet

Table 1: The intermediate result of running tag2triplet when we input the sentence  $S$  and its corresponding tag sequence  $S_{tag}$ , and we highlight the **correct** combination of candidates in bold.

$S$	... [the chronic inflammation] $_{E_0}$ ... [Helicobacter pylori infection] $_{E_1}$ ... [an increased acid production] $_{E_2}$ ...		
$S_{tag}$	[B-Emb I-Emb I-Emb] $_{E_0}$	[B-C I-C I-C ] $_{E_1}$	[B-E I-E I-E I-E] $_{E_2}$
Index	[5, 6, 7] $_{E_0}$	[17, 18, 19] $_{E_1}$	[22, 23, 24, 25] $_{E_2}$
Out-degree	1 $_{E_0}$	1 $_{E_1}$	0 $_{E_2}$
In-degree	1 $_{E_0}$	0 $_{E_1}$	1 $_{E_2}$
Candidates	$(E_0, E_2)$	$(E_1, E_0)$	$(E_1, E_2)$
Combinations	<b><math>(E_0, E_2), (E_1, E_0)</math></b>	$(E_0, E_2), (E_1, E_2)$	$(E_1, E_0), (E_1, E_2)$
Out-degree	<b><math>(1_{E_0}, 1_{E_1}, 0_{E_2})</math></b>	$(1_{E_0}, 1_{E_1}, 0_{E_2})$	$(1_{E_0}, 0_{E_1}, 0_{E_2})$
In-degree	<b><math>(1_{E_0}, 0_{E_1}, 1_{E_2})</math></b>	$(0_{E_0}, 0_{E_1}, 1_{E_2})$	$(1_{E_0}, 0_{E_1}, 1_{E_2})$

### 2.3. SCITE

Fig. 7 gives the main structure of our model SCITE for causality sequence labeling. We take the input sentence  $S = \{x_t\}_{t=1}^n$  and its corresponding label sequence  $y = \{y_i\}_{t=1}^n$  as an example to introduce each component of SCITE from bottom to top as follows, where  $n$  is the length of the  $S$ .

#### 2.3.1. CNN for Character Representations

To capture task-specific subword features, we take the same convolutional neural network [34] (CNN) architecture as Ma and Hovy [35], using a one-layer CNN structure followed by a max-over-time pooling operation [36] to learn character-level representations. The process is depicted on the left side of Fig. 7.

Specifically, let  $\mathbf{r}_i \in \mathbb{R}^m$  be the  $m$ -dimensional character vector corresponding to the  $i$ -th character in the word  $x_t$  (the length of  $x_t$  is  $s$ ). A convolution operation involves a filter  $\mathbf{w} \in \mathbb{R}^{lm}$ , which is applied to a window of  $l$  characters to produce a new feature. For example, a feature  $c_i$  is generated from a window

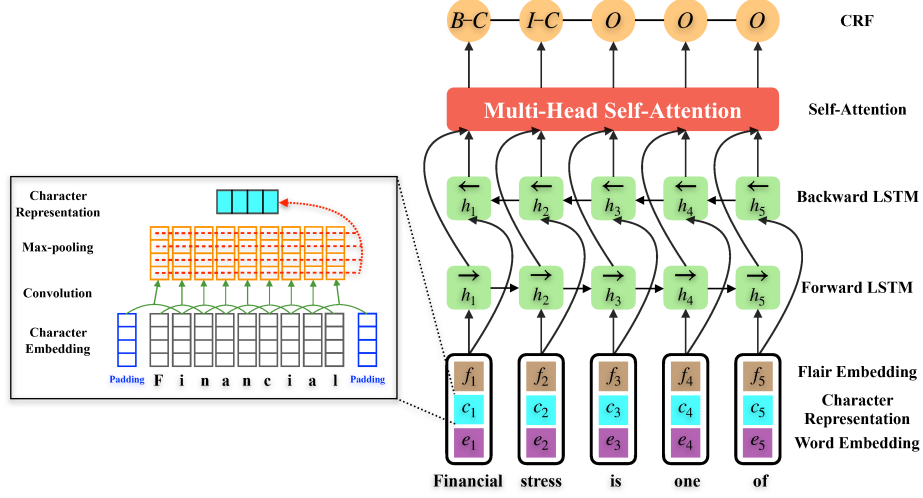


Figure 7: The main structure of SCITE for causality sequence labeling. The left side of the figure shows a character CNN structure representing the word “financial”.

of characters  $\mathbf{r}_{i:i+l-1}$ <sup>2</sup> by

$$c_i = \mathbf{w}^T \mathbf{r}_{i:i+l-1} + b, \quad (1)$$

where  $b$  is a bias term. This filter is applied to each possible window of character vectors in the word  $\{\mathbf{r}_{1:l}, \mathbf{r}_{2:h+1}, \dots, \mathbf{r}_{s-l+1:s}\}$  to produce a feature map

$$\hat{\mathbf{c}} = [c_1, c_2, \dots, c_{s-l+1}], \quad (2)$$

with  $\hat{\mathbf{c}} \in \mathbb{R}^{s-l+1}$ . Then, we take the maximum value  $\tilde{c} = \max\{\hat{\mathbf{c}}\}$  as the feature corresponding to this particular filter. Thus, denoting that the number of filters is  $f$ , the character representation  $\mathbf{c}_t$  for word  $x_t$  is given as:

$$\mathbf{c}_t = [\tilde{c}_1, \tilde{c}_2, \dots, \tilde{c}_f] \quad (3)$$

<sup>2</sup>In general, let  $\mathbf{r}_{i:i+j}$  refer to the concatenation of character vectors  $\mathbf{r}_i, \mathbf{r}_{i+1}, \dots, \mathbf{r}_{i+j}$

### 2.3.2. Transferring Contextualized Representations Learned from Large Corpus

In recent years, deep learning has ushered in incredible advances in natural language processing (NLP) tasks due to its powerful representation learning ability. However, in the case of data insufficiency of the existing corpus, the data-hungry nature of deep learning limits the performance of our neural-based model in causality extraction. The recent development of contextualized language representation models [37, 38, 31] trained on large corpora shed light on the possibility of transfer learning.

In this paper, we use transfer learning to alleviate the problem of data insufficiency. Specifically, we propose to transfer the Flair embeddings [31], which were derived from a character-level language model (CharLM) trained on a 1-billion word benchmark corpus [39] to our task. This CharLM consists of a forward language model (fLM) and a backward language model (bLM). Following Akbik et al. [31], we extract the output hidden state  $\overrightarrow{h}_{end+1}^t$  from the fLM after the last character  $r_{end}^t$  of the word  $x_t$ . Similarly, we obtain the output hidden state  $\overleftarrow{h}_{start-1}^t$  from the bLM before the first character  $r_{start}^t$  of the word  $x_t$ . Then, both output hidden states are concatenated to form the final embedding  $\mathbf{f}_t^{CharLM}$  of the word  $x_t$  as follows:

$$\mathbf{f}_t^{CharLM} = [\overrightarrow{h}_{end+1}^t, \overleftarrow{h}_{start-1}^t] \quad (4)$$

Finally, we concatenate transferred Flair embeddings  $\mathbf{f}_t^{CharLM}$  and the character representations  $\mathbf{c}_t$  with the word embeddings  $\mathbf{e}_t$  pretrained by Komninos and Manandhar [40] and feed them into a BiLSTM layer.

### 2.3.3. BiLSTM

Long short-term memory (LSTM) [41] is a particular recurrent neural network (RNN) that overcomes the vanishing and exploding gradient problems [42] of traditional RNN models. Through the specifically designed gate structure of LSTM, the model can selectively save context information. The basic unit of

the LSTM architecture is a memory block, which includes a memory cell (denoted as  $\mathbf{m}$ ) and three adaptive multiplication gates (i.e., an input gate  $\mathbf{i}$ , a forget gate  $\mathbf{f}$  and an output gate  $\mathbf{o}$ ). Formally, the computational operations to update an LSTM unit at time  $t$  are:

$$\mathbf{i}_t = \sigma(\mathbf{W}_i [\mathbf{e}_t, \mathbf{c}_t, \mathbf{f}_t^{CharLM}] + \mathbf{U}_i \mathbf{h}_{t-1} + \mathbf{b}_i), \quad (5)$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_f [\mathbf{e}_t, \mathbf{c}_t, \mathbf{f}_t^{CharLM}] + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{b}_f), \quad (6)$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_o [\mathbf{e}_t, \mathbf{c}_t, \mathbf{f}_t^{CharLM}] + \mathbf{U}_o \mathbf{h}_{t-1} + \mathbf{b}_o), \quad (7)$$

$$\widetilde{\mathbf{m}}_t = \tanh(\mathbf{W}_m [\mathbf{e}_t, \mathbf{m}_t, \mathbf{f}_t^{CharLM}] + \mathbf{U}_m \mathbf{h}_{t-1} + \mathbf{b}_m), \quad (8)$$

$$\mathbf{m}_t = \mathbf{i}_t \odot \widetilde{\mathbf{m}}_t + \mathbf{f}_t \odot \mathbf{m}_{t-1}, \quad (9)$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{m}_t), \quad (10)$$

where  $[\mathbf{e}_t, \mathbf{c}_t, \mathbf{f}_t^{CharLM}]$  and  $\mathbf{h}_t$  represent the input vector and hidden state, respectively at time  $t$ .  $\sigma$  is the elementwise sigmoid function, and  $\odot$  is the elementwise product.  $\mathbf{W}_i, \mathbf{W}_f, \mathbf{W}_o, \mathbf{W}_m$  are the weight matrices for the input vector,  $\mathbf{U}_i, \mathbf{U}_f, \mathbf{U}_o, \mathbf{U}_m$  are the weight matrices for the hidden state, and  $\mathbf{b}_i, \mathbf{b}_f, \mathbf{b}_o, \mathbf{b}_m$  denote the bias vectors.

However, LSTM only considers the information from the past, ignoring future information. To efficiently use contextual information, we can use bidirectional LSTM (BiLSTM). BiLSTM uses a forward LSTM and a backward LSTM for each sequence to obtain two separate hidden states:  $\vec{\mathbf{h}}_t, \overleftarrow{\mathbf{h}}_t$ , and then the final output at time  $t$  is formed by concatenating these two hidden states:

$$\mathbf{h}_t = [\vec{\mathbf{h}}_t, \overleftarrow{\mathbf{h}}_t] \quad (11)$$

Therefore, the final output of the BiLSTM layer for the input sentence  $S$  can be represented by  $\mathbf{H} = \{\mathbf{h}_t\}_{t=1}^n$ , where  $\mathbf{H} \in \mathbb{R}^{n \times d}$ , and  $d$  is the layer size of the BiLSTM layer.

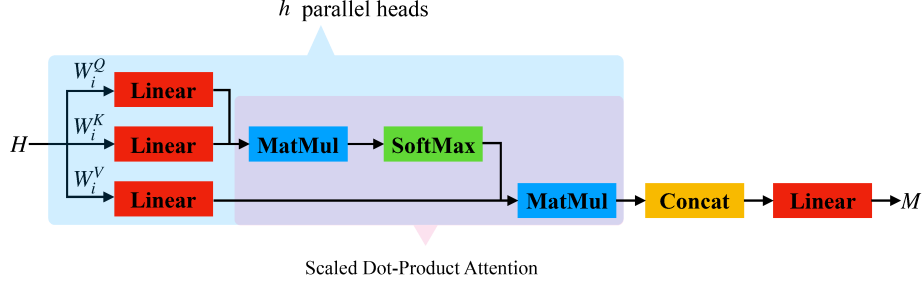


Figure 8: The architecture of the multihead attention mechanism.

#### 2.3.4. Multihead Self-Attention

Self-attention is a particular case of the attention mechanism, which only requires a single sequence to compute its representation, has been successfully applied to many NLP tasks [43, 44, 32] and shows its superiority in capturing long-range dependency. In SCITE, we adopt the multihead self-attention (MHSA) proposed by Vaswani et al. [32] to learn the dependencies of causalities in the given sentences. Fig. 8 depicts the architecture of the multihead attention mechanism.

Specifically, given  $\mathbf{H}$  as the output of the BiLSTM layer, the multihead attention mechanism first projects the matrix  $\mathbf{H}$   $h$  times with different learned linear projections to matrices:  $\mathbf{H}\mathbf{W}_i^Q$ ,  $\mathbf{H}\mathbf{W}_i^K$  and  $\mathbf{H}\mathbf{W}_i^V$ . where  $h$  is the number of heads and parameter matrices  $\mathbf{W}_i^Q \in \mathbb{R}^{d \times d_v}$ ,  $\mathbf{W}_i^K \in \mathbb{R}^{d \times d_v}$  and  $\mathbf{W}_i^V \in \mathbb{R}^{d \times d_v}$  are projections for the  $i$ -th head. Then, the attention function is performed in parallel, yielding  $n \times d_v$ -dimensional output values. Finally, all the matrices produced by parallel heads are concatenated, resulting in the final values  $\mathbf{M}$  whose dimension is  $n \times (hd_v)$ , where both  $h$  and  $d_v$  are hyperparameters of the self-attention layer. The formulations can be shown as follows:

$$\mathbf{M} = \text{MultiHead}(\mathbf{H}, \mathbf{H}, \mathbf{H}) = \text{Concat}(\mathbf{head}_1, \dots, \mathbf{head}_h) \quad (12)$$

$$\text{where } \mathbf{head}_i = \text{Attention}(\mathbf{H}\mathbf{W}_i^Q, \mathbf{H}\mathbf{W}_i^K, \mathbf{H}\mathbf{W}_i^V) \quad (13)$$

Here, the attention function is the “scaled dot-product attention”, which

computes the attention scores as follows:

$$Attention(\mathbf{H}\mathbf{W}_i^Q, \mathbf{H}\mathbf{W}_i^K, \mathbf{H}\mathbf{W}_i^V) = softmax(\frac{(\mathbf{H}\mathbf{W}_i^Q)(\mathbf{H}\mathbf{W}_i^K)^T}{\sqrt{d}})(\mathbf{H}\mathbf{W}_i^V) \quad (14)$$

To fully integrate the information, we concatenate  $\mathbf{H}$  and  $\mathbf{M}$  into matrix  $\tilde{\mathbf{H}}$  and then project  $\tilde{\mathbf{H}}$  with a linear projection to matrix:  $\tilde{\mathbf{H}}\mathbf{W}$ . where weight matrix  $\mathbf{W} \in \mathbb{R}^{(d+hd_v)k}$  is the parameter of the model to be learned in training and  $k$  is the number of distinct tags.

### 2.3.5. CRF

The conditional random field (CRF) [45] can obtain a globally optimal chain of labels for a given sequence considering the correlations between adjacent tags. In a sequence labeling task, there are usually strong dependencies between the output labels. Therefore, instead of only using RNN to model tagging decisions separately, we adopt BiLSTM-CRF [27] as the backbone of SCIF1 to jointly decode labels for the whole sentence.

We use  $\mathbf{P} \in \mathbb{R}^{n \times k}$  as the matrix of scores output by the linear layer, where  $\mathbf{P}_{ij}$  represents the score of the  $j^{th}$  label of the  $i^{th}$  word within a sentence. For the sentence  $S = \{x_t\}_{t=1}^n$  and a path of tags  $y = \{y_i\}_{i=1}^n$ , CRF gives a real-valued score as follows:

$$score(S, y) = \sum_{i=0}^n \mathbf{A}_{y_i, y_{i+1}} + \sum_{i=1}^n \mathbf{P}_{i, y_i}, \quad (15)$$

where  $\mathbf{A}$  is the transition matrix, and  $\mathbf{A}_{i,j}$  denotes the score of a transition from tag  $i$  to tag  $j$ .  $y_0$  and  $y_n$  are the special tags at the beginning and the end of a sentence, so  $\mathbf{A}$  is a square matrix of size  $k+2$ . Therefore, the probability for the label sequence  $y$  given a sentence  $S$  is:

$$p(y \mid S) = \frac{e^{score(S, y)}}{\sum_{\tilde{y} \in Y_S} e^{score(S, \tilde{y})}}, \quad (16)$$

We now maximize the log-likelihood of the correct tag sequence:

$$\log(p(y | S)) = \text{score}(S, y) - \log\left(\sum_{\tilde{y} \in Y_S} e^{\text{score}(S, \tilde{y})}\right), \quad (17)$$

where  $Y_S$  represents all possible tag sequences for an input sentence  $S$ . From the formulation above, we can obtain a valid output sequence. When decoding, the sequence with the maximum score is output by:

$$y^* = \arg \max_{\tilde{y} \in Y_S} \text{score}(S, \tilde{y}) \quad (18)$$

This can be computed using dynamic programming techniques, and we choose the Viterbi algorithm [46] for this decoding.

### 3. Experiments

#### 3.1. Experimental Settings

##### 3.1.1. Dataset

In the experiment, we evaluate a corpus obtained by extending the annotations of the SemEval 2010 task 8 dataset. [28]. In the original dataset, only one causal triplet in each sentence was annotated. We extend the annotation with the causal triplets not considered by the SemEval annotators; for example, we annotate all of the causal triplets in the sentence in Fig. 2 (more examples are shown in Fig. 3, Fig. 5 and Fig. 6). Specifically, the corpus is composed of 5,236 sentences, of which 1,270 sentences contain at least one causal triplet. The training set consists of 4,450 sentences and contains 1,570 causal triplets. There are 804 sentences in the test set, including 296 causal triplets. Table 2 shows the statistics of six types of causal tags for the dataset.

##### 3.1.2. Evaluation

We use standard precision (P), recall (R) and F1-score (F) as evaluation metrics, which can be calculated by the following formulas:



Table 2: Statistics of different types of causal tags for the dataset

Tag Type	Training Set	Test Set
B-C	1308	236
I-C	1421	229
B-E	1268	238
I-E	1230	230
B-Emb	55	9
I-Emb	55	16
Sum	5337	958

$$P = \frac{\text{\#correct extracted causal triplets}}{\text{\#extracted causal triplets}}, \quad (19)$$

$$R = \frac{\text{\#correct extracted causal triplets}}{\text{\#total causal triplets in } D}, \quad (20)$$

$$F = 2 \frac{P \cdot R}{P + R}, \quad (21)$$

where  $D$  is the set of all the sentences in the dataset and a predicted causal triplet is regarded as correct if and only if it precisely matches a labeled causal triplet. To obtain comparable and reproducible F1-scores, we follow the advice of Reimers and Gurevych [47] and conduct each experiment 5 times and then report the average results and their standard deviation, as shown in Table 3.

### 3.1.3. Hyperparameters

The model is implemented by using Keras <sup>3</sup> version 2.2.4. The 300-D word embeddings pretrained by Komninos and Manandhar [40] are employed and kept fixed during the training process. Character embeddings are randomly initialized from a uniform distribution ranging in  $[-\sqrt{\frac{3}{dim}}, +\sqrt{\frac{3}{dim}}]$ , where we set  $dim = 30$ . For the character-level CNN layer, we use a one-layer CNN with 30 filters, and the window size is 3. We use the Flair framework <sup>4</sup> to compute the

<sup>3</sup><https://github.com/keras-team/keras>

<sup>4</sup><https://github.com/zalando-research/flair>

Flair embeddings. The hidden size of LSTM is set to 256. The parameters  $h$  (the number of heads) and  $d_v$  (the size of each head) of the multihead self-attention mechanism are set to 3 and 8, respectively. We use variational dropout [48] with a dropout rate of 0.5 to regularize our network. To address the exploding gradient problem, we apply gradient normalization [49] with a threshold of 5.0 to the SCITE. The optimization method of the training process is Nadam [50] with a learning rate of 0.001, and we apply a learning rate annealing method such that if the training loss does not fall for more than 10 epochs, this method will halve the learning rate. We let the minibatch size be 16. In the experiments, we perform a grid search and 10-fold cross-validation on the training set to find the optimal hyperparameters. On the test set, we select the optimal model among all 200 epochs with the highest cross-validation F1-score.

#### 3.1.4. Baselines

For a comprehensive comparison, we compare our method against several classic causality extraction methods, which can be divided into two categories: pipeline methods and sequence tagging models based on our causality tagging scheme. The pipeline methods that we use as our baselines are as follows:

- **Rules+Bayesian:** Sorgente et al. [17] performed pattern matching to extract candidate cause-effect pairs based on a set of rules and then used a Bayesian classifier and Laplace smoothing to filter noncausal pairs.
- **CausalNet:** Luo et al. [19] proposed causal strength (CS) to measure the causal strength between any two pieces of short texts, integrating necessity causality with sufficiency causality. For comparison, we add the same cause-effect extraction module as Sorgente et al. [17] to their method. We then calculate the CS score of the candidate causal pair and compare it with the threshold  $\tau$  ( $\tau$  is a tunable hyperparameter). If  $CS(c, e) > \tau$ , we conclude that  $(c, e)$  is a causal relation; otherwise,  $(c, e)$  is an erroneously extracted pair.

The sequence tagging structure used in this paper is divided into CNN-

based models and BiLSTM-based models. For the CNN-based models [51], the baselines are as follows:

- **IDCNN-Softmax**: This model uses a deep iterated dilated CNN (ID-CNN) architecture to aggregate context from the entire text, which has better capacity than traditional CNN and faster computational speed than LSTM, and then map the output of IDCNN to predict each label independently through a softmax classifier.
- **IDCNN-CRF**: This model uses the CRF classifier to maximize the label probability of the complete sentence based on IDCNN. Compared to the softmax classifier, the CRF classifier is more appropriate for tasks with strong output label dependency.

The baselines for the BiLSTM-based models are listed as follows:

- **BiLSTM-softmax** [52]: The model consists of two parts: a BiLSTM encoder and a softmax classifier.
- **BiLSTM-CRF** [27]: A classic and popular choice for sequence labeling tasks, which consists of a BiLSTM encoder and a CRF classifier.
- **CLSTM-BiLSTM-CRF** [53]: A hierarchical BiLSTM-CRF model that uses character-based representations to implicitly capture morphological features (e.g., prefixes and suffixes) through a character LSTM encoder (CLSTM) and then concatenates the character embeddings and pretrained word embeddings as the input of BiLSTM-CRF.
- **CCNN-BiLSTM-CRF** [35]: A similar hierarchical BiLSTM-CRF model uses a character CNN encoder (CCNN) instead of a CLSTM to learn the character-level embeddings.

To further analyze the performance of Flair embeddings transferred into our task, we combine the ELMo [37] and BERT [38], two powerful contextualized word representations, into our task-specific BiLSTM-CRF architecture as the experimental baselines:

- **ELMo-BiLSTM-CRF**: An extension of BiLSTM-CRF in which Peters et al. [37] concatenate pretrained static word embeddings with the ELMo (Embeddings from Language Models) representations and take them as the input of BiLSTM-CRF.
- **BERT-BiLSTM-CRF**: A similar extension in which Devlin et al. [38] added pretrained word embeddings and the BERT (bidirectional encoder representations from transformers) representations and used them as the input of BiLSTM-CRF.
- **Flair-BiLSTM-CRF**: This model is used as a strong baseline in our work, in which Akbik et al. [31] pretrained word embeddings are concatenated with the Flair embeddings and fed it into the BiLSTM-CRF model. Note that the models using Flair embeddings have achieved the current state-of-the-art results in a range of sequence labeling tasks such as named entity recognition, chunking and part-of-speech tagging [31, 54].
- **Flair+CLSTM-BiLSTM-CRF**: A simple extension in which Akbik et al. [31] added task-trained character representations learned from a CLSTM to Flair-BiLSTM-CRF.

### 3.2. Experimental Results

The performance of different models on the causality extraction is shown in Table 3. The first part is the pipeline methods (from row 2 to row 3). The second part (row 4 to row 5) is the CNN-based sequence tagging method. The third part (row 6 to row 9) is the BiLSTM-based sequence tagging method, and the fourth part (row 10 to row 13) is the sequence tagging method using contextualized word embeddings. Our SCITE model is shown in the final part, where the first row is the result of SCITE based on the proposed tagging scheme and the second row is the result of SCITE based on the general tagging scheme.

Table 3 shows that SCITE outperforms all other models with an F1-score of 0.8455 in the test set. This demonstrates the effectiveness of our proposed

Table 3: Comparison in precision (P), recall (R), and F1-score (F) on the test set with baselines.

Model	P	R	F
CausalNet	0.6211	0.5372	0.5761
Rules-Bayesian	0.6042	0.5878	0.5959
IDCNN-softmax	0.7455±0.0142	0.7074±0.0168	0.7258±0.0105
IDCNN-CRF	0.7442±0.0225	0.7142±0.0122	0.7288±0.0160
BiLSTM-softmax	0.7744±0.0183	0.7622±0.0114	0.7682±0.0138
CLSTM-BiLSTM-CRF	0.8144±0.0284	0.7412±0.0073	0.7757±0.0107
CCNN-BiLSTM-CRF	0.8069±0.0199	0.7520±0.0227	0.7780±0.0075
BiLSTM-CRF	0.7837±0.0061	0.7932±0.0087	0.7884±0.0072
BERT-BiLSTM-CRF	0.8277±0.0058	0.8209±0.0093	0.8243±0.0049
Flair+CLSTM-BiLSTM-CRF	0.8403±0.0090	0.8284±0.0125	0.8343±0.0106
ELMo-BiLSTM-CRF	0.8361±0.0135	0.8399±0.0063	0.8379±0.0092
Flair-BiLSTM-CRF	<b>0.8414±0.0079</b>	0.8351±0.0141	0.8382±0.0092
<b>SCITE</b>			
(Flair+CCNN-BiLSTM-MHSA-CRF)	0.8333±0.0042	<b>0.8581±0.0021</b>	<b>0.8455±0.0028</b>
SCITE			
(based on general tagging scheme)	0.7609±0.0170	0.7757±0.0136	0.7682±0.0145

method. Furthermore, it also shows that the sequence tagging models are better than pipeline methods.

By comparing the performance of the sequence tagging models on the test set, we can see that the BiLSTM-based models are better than the CNN-based models. The reason for the superior performance of BiLSTM-based models may be that the LSTM layer can more efficiently capture the global word context information and learn semantic representations of causality. In addition, It also shows that the performance of the model drastically improves after feeding contextualized word representations into the BiLSTM-CRF architecture. In particular, the Flair-BiLSTM-CRF achieves the highest improvement of 6.32% over the BiLSTM-CRF compared with the ELMo and BERT (increases of 6.28% and 4.55%, respectively), which indicates that the contextualized character-level word embedding is more suitable for the task of causality extraction.

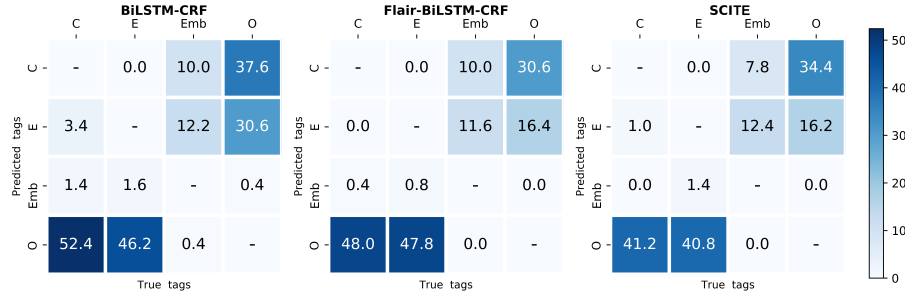


Figure 9: Confusion matrix of our SCITE model and other baseline models for tag errors. x-axis: true tags; y-axis: predicted tags.

Moreover, we also find that our proposed causality tagging scheme yields a better result than the general tagging scheme (0.8455 versus 0.7682) under the SCITE architecture, which verifies the effectiveness of our proposed tagging scheme. The general tagging scheme does not contain an “Emb” tag, and thus, the model cannot correctly identify embedded causality. Although the number of embedded causalities is relatively small in the test set, embedded causality plays a crucial role in causality extraction: one error in its identification may affect the correct extraction of multiple triplets, as shown in Fig. 3.

## 4. Analysis and Discussion

### 4.1. Error Analysis

In this paper, we focus on extracting all causal triplets from natural language texts, where the accurate identification of tags “C” (cause), “E” (effect) and “Emb” (embedded causality), which represent the semantic roles of causal events, plays a vital role in our task. To perform error analysis, we present a confusion matrix for tags “C” (including “B-C” and “I-C”), “E” (including “B-E” and “I-E”), and “Emb” (including “B-Emb” and “I-Emb”) shown in Fig. 9. We can see that most of the errors are confusion between “C”, “E” and “O”. This confusion may arise due to the problem of insufficient annotated data.

Table 4: Comparison of predicted tags concerning “C” (cause), “E” (effect) and “Emb” (embedded causality) in precision (P), recall (R), and F1-score (F) on the test set.

Model	C-P	C-R	C-F	E-P	E-R	E-F	Emb-P	Emb-R	Emb-F
BiLSTM-CRF	0.8810	0.8628	0.8718	0.8928	0.8897	0.8913	0.4343	0.0960	0.1567
Flair-BiLSTM-CRF	0.8995	0.8843	0.8917	<b>0.9294</b>	0.8885	0.9084	<b>0.8556</b>	0.1360	0.2197
<b>SCITE</b>	<b>0.8999</b>	<b>0.8998</b>	<b>0.8998</b>	0.9272	<b>0.9021</b>	<b>0.9144</b>	0.8489	<b>0.1920</b>	<b>0.2947</b>

Compared with other baselines <sup>5</sup>, our model SCITE can better identify “C”, “E”, and “Emb”.

Furthermore, we compare the tagwise performance of our SCITE model with baselines. The comparative results are summarized in Table 4. First, we observe that our model achieves No. 1 in tags “C” (including “B-C” and “I-C”), “E” (including “B-E” and “I-E”), and “Emb” (including “B-Emb” and “I-Emb”) in terms of F1-score. Second, we also notice that the F1-scores are approximately 0.9 except for tag “Emb” because of its low frequency (only 110 instances) in the training set. In particular, it can be seen from the confusion matrix in Fig. 9 that most “Emb” tags in the test set are misidentified as “C” or “E”, which leads to the low recall of “Emb”.

#### 4.2. Ablation Analysis

To investigate the effect of the different components in SCITE (Flair+CCNN-BiLSTM-MHSA-CRF), we also report the results of ablation experiments in Table 5. All parts positively contribute to the performance of the SCITE model.

Specifically, we find that the transferred Flair embeddings provide the most significant improvement. This validates our assumption that the lack of data containing causal triplets in the existing corpus will affect the performance of a neural-based model in causality extraction. Impressively, compared with SCITE without Flair embeddings (SCITE-Flair), the transferred Flair embeddings achieve an improvement of 33.28% in terms of the F1-score in the case

<sup>5</sup>For the convenience of the display, we only show the results of SICFI, Flair-BiLSTM-CRF (the superior of baselines) and BiLSTM-CRF (the classic sequence tagging model).

Table 5: Ablation analysis of our proposed model SCITE. “All” denotes the complete SCITE model, i.e., the Flair+CCNN-BiLSTM-MHSA-CRF model, while “-” denotes removing the component from the SCITE.

Model	Setting	F
<b>SCITE</b>	<b>All</b>	<b>0.8455</b>
Flair-BiLSTM-MHSA-CRF	-CCNN	0.8438
Flair-BiLSTM-CRF	-CCNN -MHSA	0.8382
BiLSTM-MHSA-CRF	-Flair -CCNN	0.8137
BiLSTM-CRF	-Flair -CCNN -MHSA	0.7884

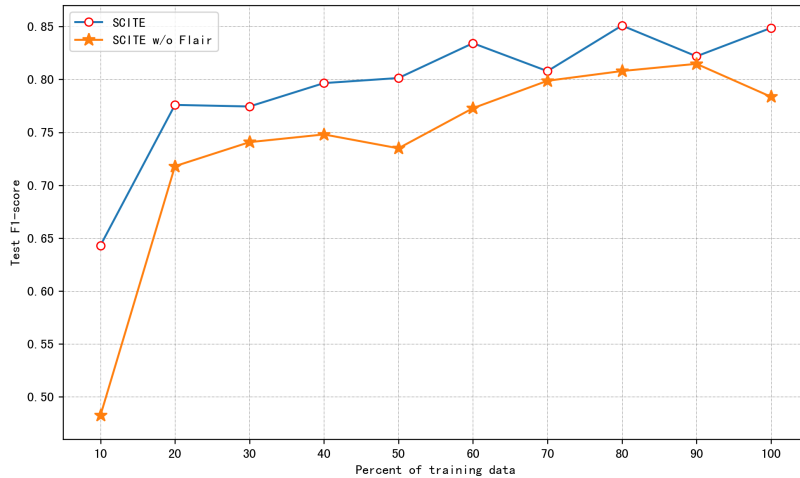


Figure 10: F1-score on the test set, in terms of the size of the training dataset.

of extremely annotated data insufficiency (10% of training data), as shown in Fig. 10. With the help of the transferred contextualized representations, we can not only learn more semantic and syntactic information from the text but also capture word meaning in context to address the polysemous and context-dependent nature of words.

In addition, we also find that the multihead self-attention (MHSA) mechanism can further improve performance, especially when there are no Flair embeddings; the reason is discussed in Section 4.3. Finally, we find that the task-specific character features can also influence the performance of the model



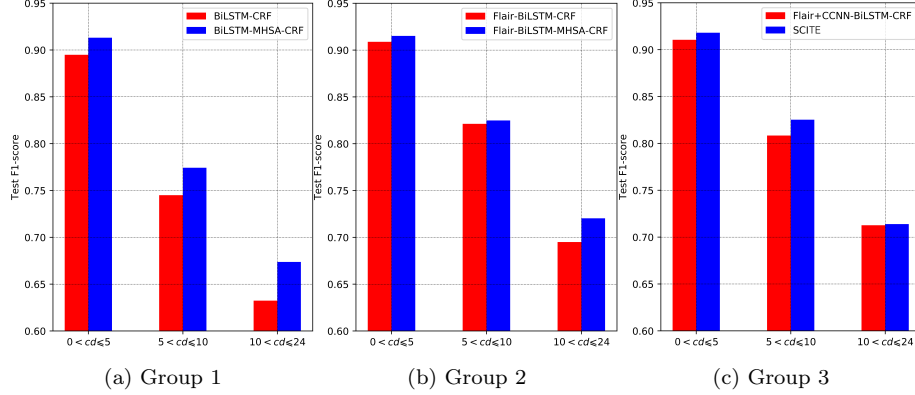


Figure 11: Comparison of the F1-score, in terms of causality distance ( $cd$ ). We divide the causal triplets in the test set into three parts according to their  $cd$ :  $0 < cd \leq 5$ ,  $5 < cd \leq 10$ , and  $10 < cd \leq 24$  (the maximum  $cd$  in the test set is 24), the ratio is approximately 2:2:1.

by a slight increase when comparing the models with and without the character representations learned from a CCNN.

#### 4.3. Analysis of Multihead Self-Attention

Different from other sequence tagging models, SCITE uses the multihead self-attention mechanism to learn the dependencies between cause and effect. To further analyze the effect of the MHSA, we compute and visualize the F1-score in terms of causality distance (the distance between cause and effect) for the three groups of models:

- Group 1: BiLSTM-CRF and BiLSTM-MHSA-CRF;
- Group 2: Flair-BiLSTM-CRF and Flair-BiLSTM-MHSA-CRF;
- Group 3: Flair+CCNN-BiLSTM-CRF and SCITE (Flair+CCNN-BiLSTM-MHSA-CRF)

As shown in Fig. 11, we find that the F1-scores decrease with increasing causality distance in all three groups. This validates our assumption that the long-range dependency between cause and effect creates difficulty in causality

extraction. In addition, we also see that the performance of models with MHSA is better than that of models without MHSA in arbitrary causality distance, which indicates that the MHSA mechanism plays a crucial role in efficiently enhancing the association between cause and effect. In particular, MHSA significantly improves the performance in terms of the causality distance greater than 10 compared with other cases of shorter causality distance, as shown in Fig. 11a and Fig. 11b.

#### 4.4. Case Study

In Table 6, we list two representative examples to show the advantages and disadvantages of our proposed model. For each case, we show the input sentence and causal triplets contained in the sentence in the first and second row. The remaining rows show the extracted causal triplets of different models <sup>5</sup>.

Sentence 1 is the case of simple causality (see Section 2.2.1), in which five causal triplets are waiting for models to extract. We observed that neither the SCITE model nor the other two baseline models could obtain all causal triplets correctly. It seems to be difficult for the models to learn that “superficial or underground water” is a complete semantic unit and the phrase “as well as” may play a key role in connecting two causal components. The reason may be that the sequence tagging models based on our causality tagging scheme require slightly more training data to learn these kinds of causality expression patterns.

Sentence 2 is the case of complex causality (see Section 2.2.2), in which there is an embedded causality and therefore brings difficulty and ambiguity to the causality learning of the model. In this example, only the SCITE can capture all the dependencies between cause and effect and thus precisely extract all three causal triplets when compared with other models.

## 5. Related Works

In this section, we briefly introduce causality extraction techniques proposed by other researchers, which fall into three categories: 1) approaches that employ

Table 6: Results of causality extraction, where “C” represents “cause”, and “E” represents “effect”. We italicize *correct* results and highlight the **incorrect** results in bold.

Sentence 1	[The damages] caused by [mudslides], [tremors], [subsidence], [superficial or underground water] were verified, as well as [swelling clay soils].
True Triplets	{[mudslides], C-E, [The damages]}, {[tremors], C-E, [The damages]}, {[subsidence], C-E, [The damages]}, {[superficial or underground water], C-E, [The damages]}, {[swelling clay soils], C-E, [The damages]}
SCITE	{[mudslides], C-E, [The damages]}, {[tremors], C-E, [The damages]}, {[subsidence], C-E, [The damages]}, <b>{[superficial], C-E, [The damages]},</b> <b>{[underground water], C-E, [The damages]}</b>
Flair-BiLSTM-CRF	{[mudslides], C-E, [The damages]}, {[tremors], C-E, [The damages]}, {[subsidence], C-E, [The damages]}, <b>{[underground water], C-E, [The damages]}</b> <b>None</b>
BiLSTM-CRF	{[mudslides], C-E, [The damages]}, {[tremors], C-E, [The damages]}, {[subsidence], C-E, [The damages]}, <b>{[underground water], C-E, [The damages]}</b> <b>None</b>
Sentence 2	This year’s Nobel Laureates in Physiology or Medicine made the remarkable and unexpected discovery that [inflammation] in the stomach as well as [ulceration] of the stomach or duodenum is the result of [an infection] of the stomach caused by [the bacterium <i>Helicobacter pylori</i> ].
True Triplets	{[an infection], C-E, [inflammation]}, {[an infection], C-E, [ulceration]}, {[the bacterium <i>Helicobacter pylori</i> ], C-E, [an infection]}
SCITE	{[an infection], C-E, [inflammation]}, {[an infection], C-E, [ulceration]}, {[the bacterium <i>Helicobacter pylori</i> ], C-E, [an infection]}
Flair-BiLSTM-CRF	<b>{[the bacterium <i>Helicobacter pylori</i>], C-E, [inflammation]},</b> <b>{[the bacterium <i>Helicobacter pylori</i>], C-E, [ulceration]},</b> {[the bacterium <i>Helicobacter pylori</i> ], C-E, [an infection]}
BiLSTM-CRF	<b>{[the bacterium <i>Helicobacter pylori</i>], C-E,</b> <b>[the remarkable and unexpected discovery]},</b> <b>None,</b> {[the bacterium <i>Helicobacter pylori</i> ], C-E, [an infection]}

pattern matching only, 2) techniques based on the combination of patterns and machine learning, and 3) methods based on deep learning techniques.

### *5.1. Pattern-Based Methods*

Pattern-based methods extract causality through pattern matching using semantic features, lexicon-syntactic features, and self-constructed constraints. For example, Khoo et al. [1] extracted causal knowledge from the Wall Street Journal using linguistic clues and pattern matching. In the domain of the medical abstract, Khoo et al. [11] used graphical patterns to extract causal knowledge from a medical database. Girju and Moldovan [15] extracted causal relations using the syntactic pattern “NP1 causal-verb NP2” with causative verbs and then employed semantic constraints to classify candidates as causal or noncausal. Ittoo and Bouma [16] proposed a causal pair extraction method based on part-of-speech, syntactic analysis, and causality templates. In their work, causality templates were first extracted using causal sentences on Wikipedia, and then they used these templates to extract causal relations in other sentences.

These methods that rely solely on rules for pattern matching often have poor cross-domain applicability and may require extensive domain knowledge in solving problems in a particular area, as well as formulating rules that consume significant amounts of time and effort.

### *5.2. Methods Based on the Combination of Patterns and Machine Learning*

Methods based on the combination of patterns and machine learning techniques mainly treat this task in a pipeline manner. They first extract candidate phrase (or entity, event) pairs that may have causal relations according to templates or some clue words and then classify the candidate causal pairs according to some statistical features or semantic features and grammatical features to filter noncausal pairs. Girju [5] used constraints based on causality trigger words to extract causal relations in English texts and used the C4.5 decision tree to perform classification. Sorgente et al. [17] used predefined templates to extract candidate causal pairs and then used Bayesian classifier and Laplace smoothing

to filter noncausal pairs. Zhao et al. [18] proposed a new feature called causal connectives by computing the similarity of the syntactic dependency structure of sentences. They run a partial parser to extract candidate noun phrases first and then classified the candidate causal pairs using the restricted hidden naive Bayes learning algorithm in combination with other features, but their method cannot discriminate the causes from the effects. Luo et al. [19] extracted cause-effect terms from large-scale web text corpora using causal cues and then used a new statistical metric-based pointwise mutual information (PMI) to measure causal strength between any two pieces of short texts.

The above methods divide causality extraction into two subtasks: candidate causal pair extraction and relation classification (filtering noncausal pairs). The results of candidate causal pair extraction may affect the performance of relation classification and generate cascading errors. These methods often require considerable human effort and time in feature engineering, relying heavily on the manual selection of textual features, and the hand-selected features are relatively too simple to capture the in-depth semantic information of the context.

### *5.3. Methods Based on Deep Learning Techniques*

Due to the powerful representation learning capabilities of deep neural networks that can effectively capture implicit and ambiguous causal relations, the adoption of deep learning techniques for causality extraction has become a popular choice for researchers in recent years. de Silva et al. [21] used CNN to classify causal relations in the text. Kruengkrai et al. [22] used multicolumn CNN with the background knowledge extracted from noisy texts to classify such common-sense causalities as “smoke cigarettes”  $\rightarrow$  “die or lung cancer”. Similarly, Li and Mao [24] proposed a knowledge-oriented CNN that incorporates prior knowledge from lexical knowledge bases for causal relation classification. Martínez-Cámara et al. [23] proposed an LSTM-based model only fed with word embeddings for the task of causality classification. In addition to classifying causality from a common sense reasoning standpoint, Dasgupta et al. [25] and Dunietz et al. [26] also identified the linguistic expressions of causality in the text from a linguistic

point of view through deep LSTM-based models.

The main differences between our proposed method and the above methods based on deep learning techniques can be summarized as follows:

- Our method aims to automatically extract such common sense causal triplets as  $c$  in the text (Fig. 1), not only to classify causal relations or to identify the linguistic expressions of causality.
- Our method can easily handle multiple causal triplets and embedded causality in the same sentence (Section 2.1 and Section 2.2) without having to divide the sentences into subsentences that contain only one instance of causality and thus generate cascading errors as in Dasgupta et al. [25].

## 6. Conclusion

In this paper, we formulate causality extraction as a sequence tagging problem and deliver a self-attentive BiLSTM-CRF-based solution for the causality extraction. In particular, we propose SCITE to extract causality in natural language text based on our causality tagging scheme. To alleviate the problem of data insufficiency, we transfer the Flair embeddings trained from a large corpus into our task. In addition, we introduce the multihead self-attention mechanism to learn the dependencies between cause and effect. Experimental results demonstrate the effectiveness of our proposed method. However, the performance of SCITE is still limited to some extent by the insufficiency of high-quality annotated data (Section 4.4).

In future work, we will attempt to solve this problem as follows:

1. Develop annotated datasets from multiple sources based on existing datasets and our causality tagging scheme.
2. Combine our method with distant supervision [55] and reinforcement learning [56] to achieve better performance without having to build a high-quality annotated corpus for causality extraction.

## 7. Acknowledgments

This research is partially supported by the National Natural Science Foundation of China (No. U1711263).

## References

## References

- [1] C. S. G. Khoo, J. Kornfilt, R. N. Oddy, S. H. Myaeng, Automatic Extraction of Cause-Effect Information from Newspaper Text Without Knowledge-based Inferencing, *Literary and Linguistic Computing* 13 (4) (1998) 177–186, doi:10.1093/lle/13.4.177.
- [2] C. S. G. Khoo, S. H. Myaeng, R. N. Oddy, Using cause-effect relations in text to improve information retrieval precision, *Information Processing & Management* 37 (1) (2001) 119–145, doi:10.1016/s0306-4573(00)00022-4.
- [3] C. Silverstein, S. Brin, R. Motwani, J. D. Ullman, Scalable Techniques for Mining Causal Structures, *Data Mining and Knowledge Discovery* 4 (2/3) (2000) 163–192, doi:10.1023/A:1009891813863.
- [4] K. Radinsky, S. Davidovich, S. Markovitch, Learning causality for news events prediction, in: *Proceedings of the 21st international conference on World Wide Web - WWW '12*, ACM Press, 909–918, doi:10.1145/2187836.2187958, 2012.
- [5] R. Girju, Automatic detection of causal relations for Question Answering, in: *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering*, Association for Computational Linguistics, 76–83, doi:10.3115/1119312.1119322, 2003.
- [6] D.-S. Chang, K.-S. Choi, Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities, *Information Processing & Management* 42 (3) (2006) 662–678, doi:10.1016/j.ipm.2005.04.004.

- [7] A. Sobrino, C. Puente, J. Olivas, Extracting answers from causal mechanisms in a medical document, *Neurocomputing* 135 (2014) 53–60, doi:10.1016/j.neucom.2013.05.056.
- [8] 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, IEEE, 361–368, doi:10.1109/isc.2010.19, 2010.
- [9] 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: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 987–997, doi:10.3115/v1/p14-1093, 2014.
- [10] E. J. M. Ackerman, Extracting a Causal Network of News Topics, in: On the Move to Meaningful Internet Systems: OTM 2012 Workshops, Springer Berlin Heidelberg, 33–42, doi:10.1007/978-3-642-33618-8\_5, 2012.
- [11] C. S. G. Khoo, S. Chan, Y. Niu, Extracting causal knowledge from a medical database using graphical patterns, in: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics - ACL '00, Association for Computational Linguistics, 336–343, doi:10.3115/1075218.1075261, 2000.
- [12] S. Zhao, M. Jiang, M. Liu, B. Qin, T. Liu, CausalTriad: Toward Pseudo Causal Relation Discovery and Hypotheses Generation from Medical Text Data, in: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB '18, ACM Press, 184–193, doi:10.1145/3233547.3233555, 2018.
- [13] Y. Ding, J. Tang, F. Guo, Identification of drug-side effect association via multiple information integration with centered kernel alignment, *Neurocomputing* 325 (2019) 211–224, doi:10.1016/j.neucom.2018.10.028.



- [14] E. C. Alemán Carreón, H. Nonaka, A. Hentona, H. Yamashiro, Measuring the influence of mere exposure effect of TV commercial adverts on purchase behavior based on machine learning prediction models, *Information Processing & Management* 56 (4) (2019) 1339–1355, doi:10.1016/j.ipm.2019.03.007.
- [15] R. Girju, D. I. Moldovan, Text Mining for Causal Relations, in: *Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference*, May 14-16, 2002, Pensacola Beach, Florida, USA, 360–364, 2002.
- [16] A. Ittoo, G. Bouma, Extracting Explicit and Implicit Causal Relations from Sparse, Domain-Specific Texts, in: *Natural Language Processing and Information Systems*, Springer Berlin Heidelberg, 52–63, doi:10.1007/978-3-642-22327-3\_6, 2011.
- [17] A. Sorgente, G. Vettigli, F. Mele, Automatic Extraction of Cause-Effect Relations in Natural Language Text, in: *Proceedings of the 7th International Workshop on Information Filtering and Retrieval co-located with the 13th Conference of the Italian Association for Artificial Intelligence (AI\*IA 2013)*, Turin, Italy, December 6, 2013., 37–48, 2013.
- [18] S. Zhao, T. Liu, S. Zhao, Y. Chen, J.-Y. Nie, Event causality extraction based on connectives analysis, *Neurocomputing* 173 (2016) 1943–1950, doi: 10.1016/j.neucom.2015.09.066.
- [19] Z. Luo, Y. Sha, K. Q. Zhu, S. Hwang, Z. Wang, Commonsense Causal Reasoning between Short Texts, in: *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR 2016*, Cape Town, South Africa, April 25-29, 2016., 421–431, 2016.
- [20] S. Zheng, F. Wang, H. Bao, Y. Hao, P. Zhou, B. Xu, Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme, in: *Proceedings*

- of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 1227–1236, doi:10.18653/v1/p17-1113, 2017.
- [21] T. N. de Silva, X. Zhibo, Z. Rui, M. Kezhi, Causal Relation Identification Using Convolutional Neural Networks and Knowledge Based Features, World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering 11 (6) (2017) 697–702.
  - [22] C. Kruengkrai, K. Torisawa, C. Hashimoto, J. Kloetzer, J. Oh, M. Tanaka, Improving Event Causality Recognition with Multiple Background Knowledge Sources Using Multi-Column Convolutional Neural Networks, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA., 3466–3473, 2017.
  - [23] E. Martínez-Cámara, V. Shwartz, I. Gurevych, I. Dagan, Neural Disambiguation of Causal Lexical Markers Based on Context, in: IWCS 2017 - 12th International Conference on Computational Semantics - Short papers, Montpellier, France, September 19 - 22, 2017, 2017.
  - [24] P. Li, K. Mao, Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts, Expert Systems with Applications 115 (2019) 512–523, doi:10.1016/j.eswa.2018.08.009.
  - [25] T. Dasgupta, R. Saha, L. Dey, A. Naskar, Automatic Extraction of Causal Relations from Text using Linguistically Informed Deep Neural Networks, in: Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, Melbourne, Australia, July 12-14, 2018, 306–316, 2018.
  - [26] J. Dunietz, J. G. Carbonell, L. S. Levin, DeepCx: A transition-based approach for shallow semantic parsing with complex constructional triggers, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, 1691–1701, 2018.

- [27] Z. Huang, W. Xu, K. Yu, Bidirectional LSTM-CRF Models for Sequence Tagging, CoRR abs/1508.01991.
- [28] I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Ó. Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, S. Szpakowicz, SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals, in: Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010, Uppsala University, Uppsala, Sweden, July 15-16, 2010, 33–38, doi:10.3115/1621969.1621986, 2010.
- [29] T. O’Gorman, K. Wright-Bettner, M. Palmer, Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation, in: Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016), Association for Computational Linguistics, 47–56, doi:10.18653/v1/w16-5706, 2016.
- [30] N. Mostafazadeh, A. Grealish, N. Chambers, J. Allen, L. Vanderwende, CaTeRS: Causal and Temporal Relation Scheme for Semantic Annotation of Event Structures, in: Proceedings of the Fourth Workshop on Events, Association for Computational Linguistics, 51–61, doi:10.18653/v1/w16-1007, 2016.
- [31] A. Akbik, D. Blythe, R. Vollgraf, Contextual String Embeddings for Sequence Labeling, in: Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018, 1638–1649, 2018.
- [32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is All you Need, in: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, 6000–6010, 2017.
- [33] Z. Tan, M. Wang, J. Xie, Y. Chen, X. Shi, Deep Semantic Role Labeling

- With Self-Attention, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, 4929–4936, 2018.
- [34] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Backpropagation Applied to Handwritten Zip Code Recognition, *Neural Computation* 1 (4) (1989) 541–551, doi:10.1162/neco.1989.1.4.541.
  - [35] X. Ma, E. Hovy, End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 1064–1074, doi:10.18653/v1/p16-1101, 2016.
  - [36] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. P. Kuksa, Natural Language Processing (Almost) from Scratch, *Journal of Machine Learning Research* 12 (2011) 2493–2537.
  - [37] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep Contextualized Word Representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, 2227–2237, doi:10.18653/v1/n18-1202, 2018.
  - [38] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *CoRR* abs/1810.04805.
  - [39] C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, T. Robinson, One billion word benchmark for measuring progress in statistical language modeling, in: INTERSPEECH 2014, 15th Annual Conference of the

International Speech Communication Association, Singapore, September 14-18, 2014, 2635–2639, 2014.

- [40] A. Komninos, S. Manandhar, Dependency Based Embeddings for Sentence Classification Tasks, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 1490–1500, doi:10.18653/v1/n16-1175, 2016.
- [41] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Computation 9 (8) (1997) 1735–1780, doi:10.1162/neco.1997.9.8.1735.
- [42] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE Transactions on Neural Networks 5 (2) (1994) 157–166, doi:10.1109/72.279181.
- [43] J. Cheng, L. Dong, M. Lapata, Long Short-Term Memory-Networks for Machine Reading, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 551–561, doi:10.18653/v1/d16-1053, 2016.
- [44] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, Y. Bengio, A Structured Self-attentive Sentence Embedding, CoRR abs/1703.03130.
- [45] J. D. Lafferty, A. McCallum, F. C. N. Pereira, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, in: Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001, 282–289, 2001.
- [46] A. J. Viterbi, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, IEEE Trans. Information Theory 13 (2) (1967) 260–269.
- [47] N. Reimers, I. Gurevych, Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging, in:

- Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 338–348, doi:10.18653/v1/d17-1035, 2017.
- [48] Y. Gal, Z. Ghahramani, A Theoretically Grounded Application of Dropout in Recurrent Neural Networks, in: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, 1019–1027, 2016.
  - [49] R. Pascanu, T. Mikolov, Y. Bengio, On the difficulty of training recurrent neural networks, in: Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013, 1310–1318, 2013.
  - [50] T. Dozat, Incorporating Nesterov Momentum into Adam, in: Proceedings of 4th International Conference on Learning Representations, Workshop Track, 2016.
  - [51] E. Strubell, P. Verga, D. Belanger, A. McCallum, Fast and Accurate Entity Recognition with Iterated Dilated Convolutions, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2670–2680, doi:10.18653/v1/d17-1283, 2017.
  - [52] P. Wang, Y. Qian, F. K. Soong, L. He, H. Zhao, Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Recurrent Neural Network, CoRR abs/1510.06168.
  - [53] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural Architectures for Named Entity Recognition, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 260–270, doi:10.18653/v1/n16-1030, 2016.

- [54] L. Borchmann, A. Gretkowski, F. Graliski, Approaching nested named entity recognition with parallel LSTM-CRFs, in: Proceedings of the PolEval 2018 Workshop, 63–73, 2018.
- [55] M. Mintz, S. Bills, R. Snow, D. Jurafsky, Distant supervision for relation extraction without labeled data, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - ACL-IJCNLP '09, Association for Computational Linguistics, 1003–1011, doi:10.3115/1690219.1690287, 2009.
- [56] R. S. Sutton, A. G. Barto, Reinforcement Learning: An Introduction, IEEE Trans. Neural Networks 9 (5) (1998) 1054–1054, doi:10.1109/TNN.1998.712192.