Incorporating Factuality Inference to Identify Document-level Event Factuality

Heng Zhang, Peifeng Li, Zhong Qian, Xiaoxu Zhu
Natural Language Processing Lab, School of Computer Science and Technology
Soochow University, Suzhou, 215006, China
zhangheng_stu@163.com, {pfli,qianzhong,xiaoxzhu}@suda.edu.cn

Abstract

Document-level Event Factuality Identification (DEFI) refers to identifying the degree of certainty that a specific event occurs in a document. Previous studies on DEFI failed to link the document-level event factuality with various sentence-level factuality values in the same document. In this paper, we innovatively propose an event factuality inference task to bridge the sentence-level and the document-level event factuality semantically. Specifically, we present a Sentence-to-Document Inference Network (SDIN) that contains a multi-layer interaction module and a gated aggregation module to integrate the above two tasks, and employ a multi-task learning framework to improve the performance of DEFI. The experimental results on the public English and Chinese DLEF datasets show that our model outperforms the SOTA baselines significantly.

1 Introduction

Document-level Event Factuality Identification (DEFI) predicts the factual property of an event from the view of a document, i.e., describing whether an event is evaluated as a fact, a counterfact, or a possibility. It is essential for many NLP applications, such as rumor detection (Qazvinian et al., 2011) and sentiment analysis (Klenner and Clematide, 2016). Based on Saurí (2008), Qian et al. (2019) summarized document-level event factuality into the following five categories: CerTain Positive (certainly happens, CT+), PoSsible Positive (possibly happens, PS+), CerTain Negative (certainly not happens, CT-), PoSsible Negative (possibly not happens, PS-), Underspecified (factuality is uncommitted, Uu). Different from Sentence-level Event Factuality Identification (SEFI) which determines the factuality based on a single sentence in which the event is located, DEFI is a more challenging task and needs to synthesize the semantics of sentence-level mentions of the document.

(S1) A Chinese city has **banned** high school students from **tearing(CT-)** up textbooks or yelling in hallways to relieve exam pressure, state media said.
(S2) In a poll on Chinese website Sina, 51% of users said they did not support the ban, they believed that **tearing(CT+)** up books can help vent emotions.
(S3) Some users commented saying that the books belonged to the students, so it was up to them what they wanted to do with it.
(S4) Others pointed out that the students were **unlikely** to **tear(PS-)** up their textbooks before their exams, and that they were simply tearing up scrap paper.
(S5) According to the Global Times, seven students in Hubei Province were allegedly expelled last year for **tearing(CT+)** up textbooks and flinging them out of the school's windows.
...
Document-level Event Factuality: **tear(CT-)**

Figure 1: An example of annotated sentence-level and document-level event factuality.

Figure 1 illustrates the relationship and differences between sentence-level and document-level event factuality. The document-level event "tear up textbooks" is denoted by the event trigger (the main word that most clearly expresses the occurrence of an event) "tear" and is mentioned in sentences S1, S2, S4, S5. In S1, since the trigger "tearing" is negated by the negative cue "banned", its sentence-level factuality is CT-. Similarly, the factuality of "tear" in S4 is PS- according to the speculative cue "unlikely". On the contrary, "tearing" in S2 and S5 are not affected by any negative or speculative information, and they have the same factuality, i.e., CT+. Although those four event mentions (the sentence containing the event trigger) have various sentence-level factuality values, from the perspective of the document, the document-level factuality value of the event "tear up textbooks" is unique, which is determined as CT- .

From Figure 1, we can observe that the factuality of a specific event at the document-level and the sentence-level semantics may be not always consistent, making it challenging when leveraging the sentence-level factuality values. All previous work did not explore the relationship between sentence-level and document-level event factuality in depth, and usually let models learn global information from the document automatically, often resulting

Findings of the Association for Computational Linguistics: ACL 2023, pages 13990–14002

in noisy semantics.

In this paper, we re-consider the document-level factuality from a new perspective, i.e., associating document-level factuality with its core event mentions that has the same factuality (e.g., S1 in Figure 1), while weakening the other event mentions with different factuality values that may impose a negative effect (e.g., S2 and S5 in Figure 1). For this purpose, we exploit the sentence-level and document-level factuality to propose a novel event factuality inference task, which can semantically bridge the above two-level factuality to determine whether an event mention is core or not. Taking Figure 1 as an example, we can assign S1, S2, and S4 with three inference labels (i.e., Support, Refute, and Conjecture), representing how they contribute to inferring the document factuality. The inference task uses the event mentions with inference labels for training, making it feasible to measure the importance of event mentions for overall factuality identification. Specifically, we propose a Sentence-to-Document Inference Network (SDIN) with two modules, where one is a multi-layer interaction module for solving the inference task and can provide useful features for DEFI through a multi-task learning framework, and the other is a gated aggregation module that can selectively aggregate semantic and factual features in event mentions for factuality identification. Overall, our main contributions can be summarized as follows:

1) We first propose an event factuality inference task to connect the sentence-level and document-level factuality, which can effectively utilize the sentence-level information and provide a new research direction for DEFI.

2) We devise a novel Sentence-to-Document Inference Network (SDIN) containing a multi-layer interaction module and a gated aggregation module for the inference and identification task, and consider a multi-task learning framework to improve the performance.

3) The experimental results on the DLEF dataset (Qian et al., 2019) demonstrate that our model achieves tremendous improvements over various strong baselines.

2 Related Work

In Sentence-level Event Factuality Identification (SEFI), Saurí (2008) and Lotan et al. (2013) used rule-based methods, and then many studies (de Marneffe et al., 2012; Saurí and Pustejovsky, 2012; Lee et al., 2015; Qian et al., 2015) used machine learning-based methods, which relied on the annotated information. Deep learning methods (He et al., 2017; Qian et al., 2018a,b; Veyseh et al., 2019) have also been widely used in recent years and have achieved significant results.

Compared with SEFI, Document-level Event Factuality Identification (DEFI) is still in the exploratory stage. Qian et al. (2019) gave the full definition of DEFI and constructed a Document-Level Event Factuality corpus (DLEF) on the basis of sentence-level annotations, in which both sentence-level and document-level factuality of event triggers were annotated. Based on the DLEF dataset, Qian et al. (2019) and Huang et al. (2019) proposed LSTM-based methods to solve the new DEFI task. Zhang et al. (2021) used cross-domain corpora to train a BERT-CRF model for detecting negation and speculation scope, and introduced scope features into DEFI. Cao et al. (2021) proposed an Uncertain Local-to-Global Network (ULGN), which integrated the local uncertainty as well as the global structure. Zhang et al. (2022b) proposed a novel Heterogeneous Semantics-Syntax-fused Network (HS^{2}N) for DEFI, which integrated semantic and syntactic information and considered inter-and-intra sentence interaction.

Recently, some work has broadened the direction of DEFI research. Qian et al. (2022a) constructed a larger DLEF-v2 corpus to expand the DLEF. A complete sentence is used in DLEF-v2 instead of a trigger word to represent an event. In addition, DLEF-v2 only annotates the document-level factuality related to the event, and no longer performs sentence-level annotations, which supports end-to-end work. Based on DLEF-v2, Qian et al. (2022a) proposed a novel end-to-end reinforced multi-granularity hierarchical attention network to learn information at different levels of granularity from tokens and sentences hierarchically. Qian et al. (2022b) proposed a new framework formulating DEFI as Machine Reading Comprehension (MRC) tasks considering both Span-Extraction and Multiple-Choice. Distinct from these mentioned efforts, Zhang et al. (2022a) constructed a new Evidence-Based Document-Level Event Factuality corpus EB-DLEF, and proposed a pipeline approach to solve the new introduced evidential sentence selection task and event factuality identification task. Zhang et al. (2023) introduced a two-stage data augmentation strategy from text to graph

Sentence Factuality	Document Factuality				
	CT+	PS+	CT-	PS-	Uu
CT+	SP	CJ	RF	CJ	UK
PS+	CJ	SP	CJ	RF	UK
CT-	RF	CJ	SP	CJ	UK
PS-	CJ	RF	CJ	SP	UK
Uu	UK	UK	UK	UK	UK

Table 1: The construction of inference labels in event factuality inference task.

via contrastive learning to solve the problem of data scarcity in DEFI.

3 Event Factuality Inference Task

3.1 Inference Labels

Inspired by claim verification (Thorne et al., 2018) and interpretability against the experimental results, we define four inference labels based on the relationship between sentence-level and document-level event factuality, which represent the local inference of the event mention on the document-level factuality. The explanation of each inference label is as follows and the constructed results are shown in Table 1.

Support (SP): The sentence-level event factuality is the same as the document-level one (both of them are NOT Uu). In this case, the event mention has a positive effect on the inference of document-level event factuality.

Refute (RF): The sentence or document has a negative factuality value (i.e., CT- or PS-), while the other does not. In this case, the event mention negatively affects the document-level factuality value.

Conjecture (CJ): The sentence or document has a speculative factuality value (i.e., PS+ or PS-, regardless of the factuality is simultaneously negative or not), while the other does not (i.e., CT+ or CT-). In this case, the event mention has a speculative impact on the inference of document-level factuality.

Unknown (UK): The sentence or document has a factuality value of Uu, regardless of whether the two values are the same. In this case, since it is difficult to determine what effect the event mention has on the inference of document-level factuality, we specify a uniform inference label of Unknown.

3.2 Task Formulation

We use $\mathbb{D}$ to represent a document and the set $\{es_i\}_{i=1}^{n}$ to denote all event mentions in the document $\mathbb{D}$. Similar to natural language inference, we treat an event mention as the hypothesis es_h and the rest of event mentions in the same document as premises $\{es_{p_i}\}_{i=1}^{m}$ ($m = n-1$). The task of event factuality inference is to classify the hypothesis es_h into the output set $\mathcal{Y}_1$ ={SP, RF, CJ, UK} based on the premises $\{es_{p_i}\}_{i=1}^{m}$, while the identification task treats $\{es_i\}_{i=1}^{n}$ as input and the output set is $\mathcal{Y}_2$ ={CT+, CT-, PS+, PS-, Uu}.

4 Methodology

We propose a Sentence-to-Document Inference Network (SDIN) for inference and identification tasks, which mainly contains two modules of multi-layer interaction and gated aggregation. The detailed model structure is shown in Figure 2.

4.1 Sentence Encoding

We feed all the event mentions into BERT (Devlin et al., 2019), using the final hidden state of `[CLS]` token to get vectors $\{\boldsymbol{s}_i\}_{i=1}^{n}(\boldsymbol{s}_i \in \mathbb{R}^d)$, which contains the hypothesis representation $\boldsymbol{s_h} \in \mathbb{R}^d$ and the set of premises $\{\boldsymbol{s_{p_i}}\}_{i=1}^{m}(\boldsymbol{s_{p_i}} \in \mathbb{R}^d)$.

4.2 Multi-layer Interaction

End Nodes and Hub Node We design end nodes and hub node as the basic unit of interaction. Before initializing the nodes, we use the following two useful features to enhance the representation of each premise es_{p_i}: 1) Inference labels: since event mentions are annotated with the sentence-level event factuality, we can assign an inference label in Table 1 to es_{p_i}, which represents the local inference of the premise with regard to the hypothesis es_h; 2) The number of inference labels: when a specific inference label appears more often in the premises, this label is more representative of the overall attitude in multiple premises towards the hypothesis, which is extremely important for classification of hypothesis.

Since both features can be expressed as discrete numbers (we use the numbers 0-3 to represent four inference labels), we denote these two features as a two-dimensional vector and convert it to a vector $\boldsymbol{s_{f_i}} \in R^d$ using a linear layer. Then we initialize the end node $\boldsymbol{e}_i^0$ by concatenating the vector $\boldsymbol{s_{p_i}}$ with its corresponding feature as follows.

$$\boldsymbol{e}_i^0 = \boldsymbol{W}_1([\boldsymbol{s_{p_i}}; \boldsymbol{s_{f_i}}]) + \boldsymbol{b}_1 \quad (1)$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{d\times 2d}$ and $\boldsymbol{b}_1 \in \mathbb{R}^d$ are trainable parameters, ";" denotes the concatenation operation.

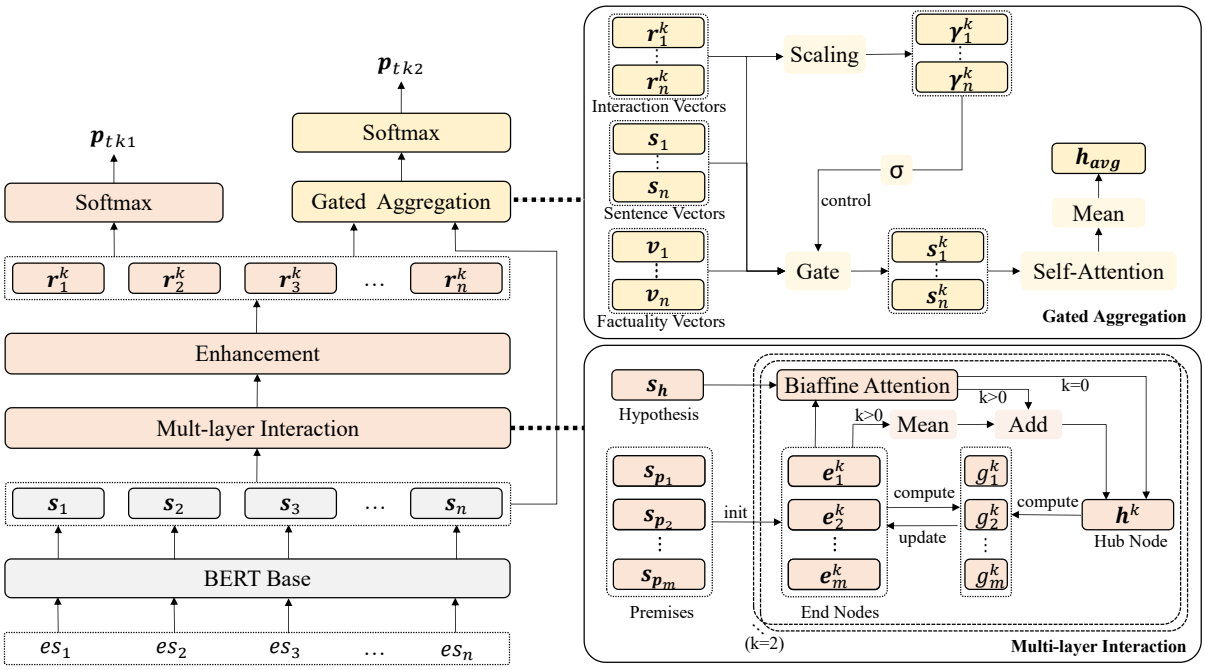

Figure 2: Overall model structure of Sentence-to-Document Inference Network (SDIN).

After obtaining the end nodes set $\boldsymbol{E} = \{\boldsymbol{e}_i^0\}_{i=1}^m$, a biaffine attention (Dozat and Manning, 2017) is used to take hypothesis as the query vector, and the result of the interactions between $\boldsymbol{s_h}$ and $\boldsymbol{E}$ is used to initialize the hub node as follows.

$$\boldsymbol{\alpha} = \text{softmax}((\boldsymbol{E}\boldsymbol{W}_2)\boldsymbol{s_h} + \boldsymbol{E}\boldsymbol{u}) \quad (2)$$

$$\boldsymbol{h}^0 = \sum_i^m \alpha_i \cdot \boldsymbol{e}_i^0 \quad (3)$$

where $\boldsymbol{W}_2 \in \mathbb{R}^{d\times d}$ and $\boldsymbol{u} \in \mathbb{R}^d$ are parameters in biaffine attention, the term $\boldsymbol{\alpha} \in \mathbb{R}^m$ is a weight vector where each element α_i in Eq. 3 denotes the coherence attention weight between the hypothesis and each end node. The result of interactions is stored in the hub node $\boldsymbol{h}^0 \in \mathbb{R}^d$.

Multi-layer Structure To make the current event mention better integrate information from the other event mentions to obtain the sentence embeddings with richer semantics, we propose a multi-layer structure that allows the end nodes and hub node to update continuously for further interactions. Each iteration starts by computing the similarity between vectors as gate values, and then we utilize the calculated results to control the update of end nodes, which can be expressed as follows.

$$g_i^k = \text{sigmoid}((\boldsymbol{h}^k)^T \boldsymbol{e}_i^k) \quad (4)$$

$$\boldsymbol{e}_i^k = g_i^{k-1} \cdot \boldsymbol{e}_i^{k-1} + (1 - g_i^{k-1}) \cdot \boldsymbol{h}^{k-1} \quad (5)$$

where the superscript k indicates the k-th layer in the stacked structure, and when k is 0, it represents the initial node. After getting the new end nodes set $\boldsymbol{E}^k = \{\boldsymbol{e}_i^k\}_{i=1}^m$, we update the hub node as follows.

$$f_{avg}(\boldsymbol{E}^k) = \frac{\sum_{i=1}^m \boldsymbol{e}_i^k}{m} \quad (6)$$

$$\boldsymbol{h}^k = \boldsymbol{\lambda} \odot f_{ba}(\boldsymbol{s_h}, \boldsymbol{E}^k) + (1 - \boldsymbol{\lambda}) \odot f_{avg}(\boldsymbol{E}^k) \quad (7)$$

where f_{ba} indicates the biaffine attention, $\odot$ means the element-wise product, $\boldsymbol{\lambda} \in \mathbb{R}^d$ is a trainable parameter that serves to obtain the interactions between hypothesis and end nodes (i.e., function f_{ba}) while not deviating from the average semantics of the end nodes (i.e., function f_{avg}).

Enhanced Representation Given the hypothesis $\boldsymbol{s_h}$ and the hub node $\boldsymbol{h}^k$ which passes through the k-layer structure, inspired by (Conneau et al., 2017; Ma et al., 2019), then we perform three matching functions and a transformation to obtain the following enhanced joint representation.

$$\boldsymbol{er}^k = [\boldsymbol{s_h}; \boldsymbol{h}^k; \boldsymbol{s_h} \odot \boldsymbol{h}^k; |\boldsymbol{s_h} - \boldsymbol{h}^k|] \quad (8)$$

$$\boldsymbol{r}^k = \tanh(\boldsymbol{W}_3 \boldsymbol{er}^k + \boldsymbol{b}_3) \quad (9)$$

where $\boldsymbol{W}_3 \in \mathbb{R}^{d\times 4d}$ and $\boldsymbol{b}_3 \in \mathbb{R}^d$ are parameters in linear layer. $\boldsymbol{r}^k$ is the final representation of the interactions of hypothesis and premises, which is used for classification in inference task and provides shared features for identification task.

4.3 Gated Aggregation

Gate Mechanism In DEFI, we take each event mention es_i as the hypothesis, and the interactive features obtained by the multi-layer interaction module is $\boldsymbol{r}_i^k$. Then, we apply a linear transformation to map $\boldsymbol{r}_i^k$ to a scaling vector, and a gate mechanism with a sigmoid function is used to generate a mask vector, which can select the most critical semantic and factual features of event mentions. The process can be formalized as follows.

$$\boldsymbol{\gamma}_i^k = \tanh(\boldsymbol{W}_4 \boldsymbol{r}_i^k + \boldsymbol{b}_4) \tag{10}$$

$$\boldsymbol{g} = [\sigma(\boldsymbol{\gamma}_i^k) \odot \boldsymbol{s}_i; \sigma(\boldsymbol{\gamma}_i^k) \odot \boldsymbol{v}_i; \sigma(\boldsymbol{\gamma}_i^k) \odot \boldsymbol{r}_i^k] \tag{11}$$

$$\boldsymbol{s}_i^k = \boldsymbol{W}_5 \boldsymbol{g} + \boldsymbol{b}_5 \tag{12}$$

where $\boldsymbol{W}_4 \in \mathbb{R}^{d \times d}$, $\boldsymbol{W}_5 \in \mathbb{R}^{d \times 3d}$, $\boldsymbol{b}_4 \in \mathbb{R}^d$, $\boldsymbol{b}_5 \in \mathbb{R}^d$ are trainable parameters, $\boldsymbol{v}_i \in \mathbb{R}^d$ is the vector of sentence-level factuality value. The result $\boldsymbol{s}_i^k \in \mathbb{R}^d$ is a more comprehensive representation of es_i, which integrates multiple features.

Interaction and Aggregation We employ the following self-attention (Vaswani et al., 2017) in set $\boldsymbol{S} = \{\boldsymbol{s}_i^k\}_{i=1}^n$ to capture the interactions of event mentions, selectively integrating information from other event mentions into the current one.

$$\boldsymbol{H} = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}})\boldsymbol{V} \tag{13}$$

$$\boldsymbol{Q} = \boldsymbol{W}_q \boldsymbol{S}, \boldsymbol{K} = \boldsymbol{W}_k \boldsymbol{S}, \boldsymbol{V} = \boldsymbol{W}_v \boldsymbol{S} \tag{14}$$

where $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, $\boldsymbol{W}_v \in \mathbb{R}^{d \times d}$ are parameters, d_k is the size of hidden units of BERT, which equals to d. Then we average each vector $\boldsymbol{h}_i \in \mathbb{R}^d$ in result $\boldsymbol{H}$ to aggregate the representation of all event mentions as follows.

$$\boldsymbol{h}_{avg} = \frac{\sum_{i=1}^n \boldsymbol{h}_i}{n} \tag{15}$$

4.4 Prediction and Joint Training

We apply the following softmax layers to $\boldsymbol{r}^k$ (Eq. 9) and $\boldsymbol{h}_{avg}$ (Eq. 15) for classification of inference and identification tasks, respectively.

$$\boldsymbol{p}_{tk1} = \text{softmax}(\boldsymbol{W}_{tk1} \boldsymbol{r}^k + \boldsymbol{b}_{tk1}) \tag{16}$$

$$\boldsymbol{p}_{tk2} = \text{softmax}(\boldsymbol{W}_{tk2} \boldsymbol{h}_{avg} + \boldsymbol{b}_{tk2}) \tag{17}$$

where $\boldsymbol{W}_{tk1} \in \mathbb{R}^{c_1 \times d}$, $\boldsymbol{W}_{tk2} \in \mathbb{R}^{c_2 \times d}$, $\boldsymbol{b}_{tk1} \in \mathbb{R}^{c_1}$, $\boldsymbol{b}_{tk2} \in \mathbb{R}^{c_2}$ are weights and biases. The cross-entropy loss for the two tasks are as follows, where N_1, N_2 are the number of samples, $\boldsymbol{y}_{tk1}^i$, $\boldsymbol{y}_{tk1}^i$ are the one-hot vector label of the i-th instance.

$$\mathcal{L}_{tk1} = -\frac{1}{N_1} \sum_{i=1}^{N_1} \boldsymbol{y}_{tk1}^i \cdot \log \boldsymbol{p}_{tk1}^i \tag{18}$$

$$\mathcal{L}_{tk2} = -\frac{1}{N_2} \sum_{i=1}^{N_2} \boldsymbol{y}_{tk2}^i \cdot \log \boldsymbol{p}_{tk2}^i \tag{19}$$

With the consideration of multi-task learning, the overall training loss is represented as follows.

$$\mathcal{L} = \mathcal{L}_{tk1} + \mathcal{L}_{tk2} \tag{20}$$

To ensure that we jointly train the two coupled tasks with intensive knowledge communication, we set the following configurations:

1) The same pre-trained model is used for both tasks with the parameters shared, which allows the sentence embeddings to be fine-tuned; 2) The two tasks share the multi-layer interaction module. On one hand, as the training of the inference task proceeds, the generated representations can provide useful information for factuality identification. On the other hand, the parameters in the multi-layer interaction will also be optimized during the training of identification task, and the noise generated by a different task will enable this module to have better generalization performance for inference; 3) The documents used for training on the inference task are never used for testing on the other task, i.e., the gold inference labels will not be obtained in advance during the testing of identification task.

5 Experimentation

5.1 Datasets and Sampling Strategy

Our experiments are conducted on the DLEF corpus (Qian et al., 2019) and the distribution of sentence-level and document-level factuality values is shown in Table 2.

We can observe that the distribution of factuality is unbalanced both in documents and sentences. The non-uniform distribution of factuality values also leads to a similar distribution of inference labels. To address this issue, we adopt the following sampling strategy to selectively choose a portion of samples for training:

1) If there is only one type of sentence-level factuality value in a document, resulting in only one inference label. In this case, we randomly select one of the event mentions as the hypothesis and add it to the training set.

Corpus	Factuality	Document	Sentence
English	CT+	1150	4401
	PS+	274	574
	CT-	279	662
	PS-	12	37
	Uu	12	71
Chinese	CT+	2403	11482
	PS+	848	2879
	CT-	1342	3923
	PS-	36	123
	Uu	20	555

Table 2: Statistics of document-level and sentence-level factuality values in the DLEF corpus.

Type	Dataset	SP	RF	CJ	UK
Training	English	1038	203	314	64
	Chinese	2364	981	1186	411
Testing	English	962	44	113	18
	Chinese	3008	245	283	140

Table 3: Inference label distribution in the training set and testing set for a random fold in the event factuality inference task.

2) If a document has more than one sentence-level factuality value, we select all the event mentions as the hypothesis that are not labeled as SP (Support), and ignore the others with SP label.

After sampling, the inference label distribution is shown in Table 3.

5.2 Experimental Settings

For a fair comparison, we perform 10-fold cross-validation on both English and Chinese corpora. Since PS- and Uu documents only cover 1.39% and 1.20% in the English and Chinese corpora, respectively, we only focus on the performance of CT+, PS+ and CT- following previous work (Qian et al., 2019; Cao et al., 2021). In addition to using F1-Score to evaluate the performance in each category, we also use Macro- and Micro-averaging F1-Score to measure the overall performance.

In our implementations, we use HuggingFace's Transformers[1] to implement the BERT base model, which has 12 layers and the hidden units d is 768. In both sub-corpora and all tasks, the optimal layers k is 2, c_1 and c_2 are equal to 4 and 5, the learning rate is 1e-5, the batch size is set to 8, and the Adam algorithm is used to optimize the model parameters.

[1]https://github.com/huggingface/transformers

5.3 Baselines

To verify the effectiveness of the proposed SDIN model, we conduct the following strong baselines for fair comparison.

1) **Att-Adv** (Qian et al., 2019): A LSTM-based model which utilizes intra- and inter-sentence attention to learn a document representation.

2) **BiLSTM** (Huang et al., 2019): A BiLSTM-based model which utilizes a double-layer attention mechanism to capture the latent correlation features among event sequences to identify the factuality.

3) **BERT-MSF** (Zhang et al., 2021): A BERT-based model which uses BERT-CRF and cross-domain corpora to detect event-related negation and speculation scope for factuality identification.

4) **ULGN** (Cao et al., 2021): A uncertain local-to-global network which models the uncertainty of local information and leverages the global structure to identify the event factuality.

5) **HS^{2}N** (Zhang et al., 2022b): A heterogeneous semantics-syntax-fused network which integrates both semantic and syntactic information, and considers both inter-and-intra sentence interaction.

6) **CoDE** (Zhang et al., 2023): A model proposes a two-stage data augmentation strategy from text to graph via contrastive learning for identifying the document-level factuality.

5.4 Overall Results

Table 4 shows the performance of each model on the DLEF corpus, from which we can draw the following conclusions:

1) Our SDIN model outperforms the baselines in all metrics. Taking Micro-F1 as an example, compared with SOTA CoDE, our model achieves 3.66 and 2.37 improvements on the English and Chinese corpora, respectively, proving its effectiveness.

2) Since the number of documents annotated in the Chinese corpus and the proportion of documents with speculative or negative factuality values are significantly larger than those in the English corpus, all models perform better in the Chinese corpus, especially in the PS+ and CT- factuality categories, where the lead is more pronounced.

3) SDIN leads all baselines by a larger margin in the much smaller DLEF English corpus, demonstrating that SDIN's relatively complex model structure can instead be better adapted to scenarios with fewer samples.

4) Pre-trained models are critical for DEFI. The methods (i.e., BERT-MSF, ULGN, HS^{2}N, CoDE,

Dataset	Method	CT+	PS+	CT-	Macro-F1	Micro-F1
DLEF English	Att-Adv	89.84	62.14	76.87	76.28	83.56
	BiLSTM	90.74	75.75	78.82	82.28	86.51
	BERT-MSF	92.50	76.38	83.71	84.24	88.64
	ULGN	92.49	76.68	84.87	84.68	88.69
	HS^2N	93.39	84.37	88.46	88.74	90.96
	CoDE	93.71	84.21	86.79	88.32	91.23
	SDIN(Ours)	**97.88**	**86.93**	**90.94**	**92.09**	**94.89**
DLEF Chinese	Att-Adv	87.52	74.06	83.35	81.64	84.03
	BiLSTM	89.74	78.52	86.09	85.05	86.64
	BERT-MSF	92.09	85.71	90.08	89.35	90.34
	ULGN	93.53	90.76	94.99	93.09	93.77
	HS^2N	92.89	88.93	94.42	92.08	92.95
	CoDE	94.26	89.53	94.96	92.92	93.77
	SDIN(Ours)	**97.48**	**92.64**	**95.95**	**95.37**	**96.14**

Table 4: Experimental results on the English and Chinese DLEF corpora.

Method	SP	RF	CJ	UK
HAN(EN)	89.10	50.70	54.51	83.19
MLA(EN)	91.23	53.79	56.88	**87.01**
SDIN(EN)	**91.80**	**55.87**	**57.47**	86.23
HAN(CN)	88.76	58.44	60.33	91.98
MLA(CN)	**90.85**	61.89	62.97	93.66
SDIN(CN)	90.71	**62.32**	**64.69**	**94.90**

Table 5: Experimental results of event factuality inference task.

SDIN) using pre-trained models can achieve significant improvements in comparison with other methods (i.e., Att-Adv, BiLSTM).

5.5 Results of Event Factuality Inference

We report the performance of event factuality inference in Table 5. The models HAN (Ma et al., 2019) and MLA (Kruengkrai et al., 2021) used for comparison are two widely used baselines in claim verification, which have a similar form to our inference task. From the experimental results we can learn that:

1) Since the SP (Support) category in the training set of the inference task has a significantly larger number of samples than the RF (Refute) and CJ (Conjecture) categories, it performs significantly better in both the Chinese and English corpora. Although the UK (Unknown) category achieves good performance with a small sample size, it shows more fluctuations during training.

2) In terms of overall performance, our SDIN model is significantly better than HAN and close to MLA. It is worth noting that we constructed this inference task to provide useful features for DEFI before pursuing better performance.

5.6 Results of End-to-end Setting

The fact that SDIN uses more sentence-level annotation information may lead to unfair comparisons. To verify the performance of SDIN using predicted information instead of gold annotations, we use the following two simple unsupervised approaches to obtain predicted event mentions and their sentence-level factuality, transforming SDIN into an end-to-end model and comparing it fairly with RMHAN (Qian et al., 2022a) and Ext-TL (Qian et al., 2022b) on the DLEF-v2 corpus (Qian et al., 2022a).

1) Event mentions: we measure the similarity between candidate sentences and events by calculating the ROUGE scores of event sentences and the remaining sentences. Specifically, the ROUGE-1 of all sentences and labeled events are calculated, sorted in reverse order based on their F1 scores, and the five sentences with the highest scores are selected as the set of event mentions.

2) Sentence-level factuality: for the event mentions obtained in 1), the factuality value is approximated according to whether they contain negative or speculative cues from BioScope (Vincze et al., 2008) and CNeSp (Zou et al., 2015) for English and Chinese, respectively. Specifically, if the event mention contains only negative/speculative cues, its factuality is CT-/PS+; If the event mention contains both negative and speculative cues, its factuality is PS-; If the event mention has the word (e.g., if,

even though) that leads to underspecified semantics, its factuality is Uu; If the event mention does not meet the above conditions, its factuality is CT+.

The above outputs are then used as an alternative to annotations to verify the performance of SDIN on the DLEF-v2 corpus, and its comparison with RMHAN and Ext-TL is shown in Table 6, from which the following conclusions can be drawn:

1) Although the DLEF-v2 corpus is significantly larger than DLEF, after applying simple unsupervised approaches for end-to-end experiments, the performance of SDIN shows a significant degradation compared to using accurate manually annotated information.

2) SDIN is still significantly leading all strong baselines. Taking Micro-F1 as an example, compared with the SOTA Ext-TL, SDIN achieves 5.43 and 3.27 improvements in the English and Chinese corpora, respectively. This proves the effectiveness of the SDIN model, and also shows that SDIN is highly robust and can better handle the large amount of noise present in the input.

5.7 Ablation Study

We conduct the following ablation studies based on the subsections of §4.2, §4.3, and §4.4 to demonstrate the effectiveness of the components in our proposed model SDIN: 1) **w/o EH**, which removes the initialization methods of end nodes and hub node, directly use the representation of the premises and their element-wise addition as the end nodes and the hub node; 2) **w/o MS**, which removes the multi-layer structure and directly use the initialized hub node as the result of the interaction; 3) **w/o ER**, which removes the enhanced representation and directly uses the hub node as the output; 4) **w/o GM**, which removes the gate mechanism and replaces it with a simple concatenation; 5) **w/o IA**, which removes the interaction and aggregation component and replaces it with an element-wise addition in all event mentions; 6) **w/o JT**, which removes the joint training approach, i.e., the multi-layer interaction module is not acted as a shared layer and the two tasks are trained in a pipeline manner. The results are presented in Table 7, and we can find that:

1) Removing EH (w/o EH), MS (w/o MS) and IA (w/o IA) have a significant impact on performance, especially in the English corpus, proving that they are more effective for factuality identification. Furthermore, the results show that capturing the interactions between event mentions is extremely critical.

2) Removing the joint training mechanism (w/o JT) will lead to a large magnitude of performance degradation, indicating that shared layers and multi-task learning are very effective for DEFI.

3) Removing the components ER (w/o ER) and GM (w/o GM) has a relatively small impact on the results as they are simply a more efficient practice for capturing some critical information, while some simple alternatives in the ablation study can also yield good results in most cases.

[D1] Document-level factuality: **affect(CT-)** **[Predict: CT-]**
(S1)...adding that the meeting will **affect(CT+)** ties between Beijing and Moscow... **[refute/refute/0.20]**
(S2)...China_US ties will not **affect(CT-)** Beijing_Moscow relations... **[support/conjecture/0.62]**
(S3)...The old relationship greatly **affects(CT+)** the existing one... **[refute/refute/0.18]**

[D2] Document-level factuality: **change(CT-)** **[Predict: CT+]**
(S1)...News outlets report that Beijing's position has **changed(CT+)**... **[refute/refute/0.33]**
(S2)...The Dalai Lama's actions won't **change(CT-)** Beijing's position on... **[support/support/0.31]**
(S3)...analysts speculate that Beijing's **change(PS+)** of position will become a foregone conclusion... **[conjecture/conjecture/0.36]**

Figure 3: Examples for case study. Taking S1 in D1 as an example, refute/refute means the true label and predicted label, and 0.20 is the calculated probability score, which is obtained by passing the outputs in Eq.10 through a sigmoid function and then taking their average. To make the results more intuitive, we normalize all the scores in a document so that the sum of the scores is 1.

5.8 Impact of Event Factuality Inference

We summarize the following three aspects to illustrate how event factuality inference can be beneficial to DEFI, just as exemplified in Figure 3.

1) All inference labels of event mentions are identified correctly. If it is true, each sentence is given an appropriate score (see the D1 in Figure 3, the scores of S1 and S3 are 0.20 and 0.18, respectively). The scores determine how much information is incorporated into the final representation, then we can get the correct factuality in complex scenarios, e.g., there are multiple different values of sentence-level factuality.

2) The inference labels of a small part of the event mentions are misidentified. In this case, if the misclassified sentence is not a core event mention, then it will not have a significant impact on final identification. Furthermore, we find that even the misclassified core event mentions tend to have

Dataset	Method	CT+	PS+	CT-	Macro-F1	Micro-F1
DLEF-v2 English	RMHAN	84.35	55.13	56.43	65.30	76.38
	Ext-TL	85.23	58.91	61.85	68.66	78.09
	SDIN(Ours)	**89.56**	**60.67**	**70.43**	**73.88**	**83.52**
DLEF-v2 Chinese	RMHAN	82.60	65.55	73.83	73.99	77.07
	Ext-TL	84.91	69.92	77.20	77.34	79.43
	SDIN(Ours)	**85.16**	**78.25**	**81.49**	**81.79**	**82.70**

Table 6: End-to-end performance on the English and Chinese DLEF-v2 corpora.

Method	DLEF English		DLEF Chinese	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1
w/o EH	89.37 (↓ 2.72)	92.81 (↓ 2.08)	93.10 (↓ 2.27)	94.72 (↓ 1.42)
w/o MS	89.66 (↓ 2.43)	92.64 (↓ 2.25)	93.38 (↓ 1.99)	94.74 (↓ 1.40)
w/o ER	90.71 (↓ 1.38)	93.43 (↓ 1.46)	94.21 (↓ 1.16)	95.44 (↓ 0.70)
w/o GM	91.03 (↓ 1.06)	93.51 (↓ 1.38)	94.07 (↓ 1.30)	95.22 (↓ 0.92)
w/o IA	89.01 (↓ 3.08)	92.27 (↓ 2.62)	93.19 (↓ 2.18)	94.05 (↓ 2.09)
w/o JT	89.91 (↓ 2.18)	92.57 (↓ 2.32)	92.92 (↓ 2.45)	93.84 (↓ 2.30)

Table 7: Results of ablation study on DLEF corpus, where the drop (↓) represents the decrease.

a positive effect through the calculated scores (e.g., S2 in D1 is misclassified, it still has a score of 0.62), and the remaining correctly classified non-core sentences have little effect on the results (e.g., S1 and S3 in D1 have low scores), so that the correct factuality can also be obtained in some cases.

3) Most of the event mentions are misclassified, often leading to errors at this point. This situation occurs infrequently and other models also fail to classify it correctly.

5.9 Error Analysis

By integrating the performance of the inference task and the example D2 in Figure 3, we summarize the following two types of errors:

1) From Table 5, we can observe that "Refute" and "Conjecture" categories perform worse than "Support", making it difficult to classify the documents that needs to obtain negative or speculative information from event mentions. Specifically, some critical event mentions are not classified correctly in inference task and obtain lower scores (the definition of scores is mentioned in Figure 3), which leads to the fact that the generated features of the critical event mentions do not have a sufficiently positive impact on DEFI.

2) In some samples, only one core event mention is valuable for prediction, and other mentions with various noisy factuality values will interfere with the correct classification. In this case, we note that even if each event mention is correctly classified in the inference task, it is still difficult for the identification model to correctly capture the most critical information from the complex relationships to infer the correct document-level factuality value. Taking D2 in Figure 3 as an example, the model correctly identifies the inference labels of S1-S3, but the core event mention S2 only obtains a score of 0.31, while the scores of S1 and S3 are 0.33 and 0.36, which lead to a representation for DEFI that does not incorporate enough core information and then leads to an error.

6 Conclusion

In this paper, we innovatively employ the relationship between the sentence- and document-level event factuality to construct an event factuality inference paradigm, which can bridge these two-level factuality semantically. Moreover, we propose a Sentence-to-Document Inference Network (SDIN), which is a multi-task learning framework incorporating a multi-layer interaction module and a gated aggregation module to integrate the task of document-level event factuality inference and identification. Experimental results on the English and Chinese DLEF datasets demonstrate the significant improvements of our proposed model SDIN, in comparison with the SOTA baselines. In future work, we will focus on end-to-end document-level event factuality identification.

Limitations

The limitations can be illustrated from the perspective of task development: DEFI originally evolved from sentence-level work, as is clearly evident from the large number of sentence-related annotations retained in DLEF corpus. Our work benefits from these abundant annotations and achieves huge performance improvements. Currently, there is a trend to gradually move towards end-to-end practice in event factuality identification. For example, the studies based on the DLEF-v2 (Qian et al., 2022a,b) and EB-DLEF (Zhang et al., 2022a) corpora have attempted to use less annotation information. Although these efforts do not achieve competitive performance for the time being, it is an exciting research direction because it allows models to be more easily applied directly to realistic scenarios. The limitation of our work lies in the fact that it runs counter to the end-to-end concept, so we need more other work (e.g, event extraction and SEFI models) to apply the model to the real world, which makes our work less applicable.

Ethics Statement

We comply with the ACL Ethics Policy. First, document-level event factuality identification is a fundamental research of natural language processing that benefits many downstream tasks. It reveals the factual property of an event in the document and does not generate any uncontrolled biased or toxic text. Second, the source data we collect come from open sources, and everyone can access them. Finally, we guarantee that the paper is completely written by authors and not automatically generated by AI models to ensure authenticity.

Acknowledgments

The authors would like to thank the three anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China (Nos. 61836007, and 62006167.), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

References

Pengfei Cao, Yubo Chen, Yuqing Yang, Kang Liu, and Jun Zhao. 2021. Uncertain local-to-global networks for document-level event factuality identification. In *EMNLP*, pages 2636–2645. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*, pages 670–680. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Comput. Linguistics*, 38(2):301–333.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *ICLR (Poster)*. OpenReview.net.

Tianxiong He, Peifeng Li, and Qiaoming Zhu. 2017. Identifying chinese event factuality with convolutional neural networks. In *CLSW*, volume 10709 of *Lecture Notes in Computer Science*, pages 284–292. Springer.

Rongtao Huang, Bowei Zou, Hongling Wang, Peifeng Li, and Guodong Zhou. 2019. Event factuality detection in discourse. In *NLPCC*, volume 11839 of *Lecture Notes in Computer Science*, pages 404–414. Springer.

Manfred Klenner and Simon Clematide. 2016. How factuality determines sentiment inferences. In **SEM@ACL*. The *SEM 2016 Organizing Committee.

Canasai Kruengkrai, Junichi Yamagishi, and Xin Wang. 2021. A multi-level attention model for evidence-based fact checking. In *ACL/IJCNLP (Findings)*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2447–2460. Association for Computational Linguistics.

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *EMNLP*, pages 1643–1648. The Association for Computational Linguistics.

Amnon Lotan, Asher Stern, and Ido Dagan. 2013. Truthteller: Annotating predicate truth. In *HLT-NAACL*, pages 752–757. The Association for Computational Linguistics.

Jing Ma, Wei Gao, Shafiq R. Joty, and Kam-Fai Wong. 2019. Sentence-level evidence embedding for claim verification with hierarchical attention networks. In *ACL*, pages 2561–2571. Association for Computational Linguistics.

Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *EMNLP*, pages 1589–1599. ACL.

Zhong Qian, Peifeng Li, Yue Zhang, Guodong Zhou, and Qiaoming Zhu. 2018a. Event factuality identification via generative adversarial networks with auxiliary classification. In *IJCAI*, pages 4293–4300. ijcai.org.

Zhong Qian, Peifeng Li, Guodong Zhou, and Qiaoming Zhu. 2018b. Event factuality identification via hybrid neural networks. In *ICONIP*, volume 11305 of *Lecture Notes in Computer Science*, pages 335–347. Springer.

Zhong Qian, Peifeng Li, and Qiaoming Zhu. 2015. A two-step approach for event factuality identification. In *IALP*, pages 103–106. IEEE.

Zhong Qian, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2019. Document-level event factuality identification via adversarial neural network. In *NAACL-HLT*, pages 2799–2809. Association for Computational Linguistics.

Zhong Qian, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2022a. Document-level event factuality identification via reinforced multi-granularity hierarchical attention networks. In *IJCAI*, pages 4338–4345. ijcai.org.

Zhong Qian, Heng Zhang, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2022b. Document-level event factuality identification via machine reading comprehension frameworks with transfer learning. In *COLING*, pages 2622–2632. International Committee on Computational Linguistics.

Roser Saurí. 2008. *A factuality profiler for eventualities in text*. Brandeis University.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? assessing the factuality degree of events in text. *Comput. Linguistics*, 38(2):261–299.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. In *NAACL-HLT*, pages 809–819. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Amir Pouran Ben Veyseh, Thien Huu Nguyen, and Dejing Dou. 2019. Graph based neural networks for event factuality prediction using syntactic and semantic structures. In *ACL*, pages 4393–4399. Association for Computational Linguistics.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinform.*, 9(S-11).

Heng Zhang, Zhong Qian, Peifeng Li, and Xiaoxu Zhu. 2022a. Evidence-based document-level event factuality identification. In *PRICAI*, volume 13630 of *Lecture Notes in Computer Science*, pages 240–254. Springer.

Heng Zhang, Zhong Qian, Xiaoxu Zhu, and Peifeng Li. 2021. Document-level event factuality identification using negation and speculation scope. In *ICONIP*, volume 13108 of *Lecture Notes in Computer Science*, pages 414–425. Springer.

Zihao Zhang, Chengwei Liu, Zhong Qian, Xiaoxu Zhu, and Peifeng Li. 2022b. Hs^2n: Heterogeneous semantics-syntax fusion network for document-level event factuality identification. In *PRICAI*, volume 13630 of *Lecture Notes in Computer Science*, pages 309–320. Springer.

Zihao Zhang, Zhong Qian, Xiaoxu Zhu, and Peifeng Li. 2023. Code: Contrastive learning method for document-level event factuality identification. In *DASFAA*, volume 13945 of *Lecture Notes in Computer Science*, pages 497–512. Springer.

Bowei Zou, Qiaoming Zhu, and Guodong Zhou. 2015. Negation and speculation identification in chinese language. In *ACL*, pages 656–665. The Association for Computer Linguistics.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section Limitations (below Section 6).

☑ A2. Did you discuss any potential risks of your work?
Section Ethics Statement (below Section Limitations).

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section Abstract and Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4.2.

☑ B1. Did you cite the creators of artifacts you used?
Section 4.2.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The artifacts we use are open source, with liberal licenses that allow and encourage NLP research.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. The artifacts we use are designed to perform specific functions and cannot be "inconsistent with their intended use". This question is not applicable to us.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We used the open source dataset and the above details have been discussed previously by the original authors.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We do not create any artifacts and the artifacts we use have detailed documentation.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4.1 and Section 4.2.

C ☑ Did you run computational experiments?

Section 4.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
The number of parameters and the computational budget depends mainly on the BERT, whose relevant information is well known by researchers. Moreover, we also provide implemented code that can be easily run on any infrastructure.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.2.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4.2.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4.2.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.