Uncertain Local-to-Global Networks for Document-Level Event Factuality Identification

Pengfei Cao[1,2]*, **Yubo Chen**[1,2]*, **Yuqing Yang**[1,3], **Kang Liu**[1,2] and **Jun Zhao**[1,2]
[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]School of Computer Science, Fudan University
{pengfei.cao, yubo.chen, kliu, jzhao}@nlpr.ia.ac.cn, yuqingyang21@m.fudan.edu.cn

Abstract

Event factuality indicates the degree of certainty about whether an event occurs in the real world. Existing studies mainly focus on identifying event factuality at sentence level, which easily leads to conflicts between different mentions of the same event. To this end, we study the problem of *document-level event factuality identification*, which determines the event factuality from the view of a document. For this task, we need to consider two important characteristics: *Local Uncertainty* and *Global Structure*, which can be utilized to improve performance. In this paper, we propose an **U**ncertain **L**ocal-to-**G**lobal **N**etwork (ULGN) to make use of these two characteristics. Specifically, we devise a *Local Uncertainty Estimation* module to model the uncertainty of local information. Moreover, we propose an *Uncertain Information Aggregation* module to leverage the global structure for integrating the local information. Experimental results demonstrate the effectiveness of our proposed method, outperforming the previous state-of-the-art model by 8.4% and 11.45% of F1 score on two widely used datasets.

1 Introduction

Event factuality refers to the degree of certainty about whether events actually occur or not in the real world. Generally, event factuality can be classified into five categories (Saurı, 2008): *Certain Positive* (certainly happening, denoted as CT+), *Certain Negative* (certainly not happening, CT-), *Possible Positive* (possibly happening, PS+), *Possible Negative* (possibly not happening, PS-) and *Underspecified* (events' factuality cannot be identified, Uu). For example, in the sentence "*An economist thinks that the tax rate probably* ***increases*** *soon*", the event "*increases*" may happen. Therefore, an event factuality identification (EFI) model should be able to predict the factuality of the event is PS+. EFI is an important task in natural language processing (NLP) area, which is beneficial for a wide range of NLP applications, such as rumor detection (Qazvinian et al., 2011), sentiment analysis (Klenner and Clematide, 2016) and machine reading comprehension (Richardson et al., 2013).

Event: United States ***reaches*** an agreement with Mexico	
	Text:
[S1]	According to Politico.com, the United States probably ***reaches*** (**PS+**) an agreement with Mexico on the new trade deal before December, 2017.
[S2]	A journalist agreed the view, said the two sides *may* ***reach*** (**PS+**) an agreement within hours.
[S3]	However, Mexican Economy Minister Ildefonso Guajardo *denied* that they plan to ***reach*** (**CT-**) any agreement with the U.S. on the trade deal talks.
[S4]	The government has not been informed that any agreement will be ***reached*** (**CT-**) yet, said another two Mexican officials.
…	
[S8]	Some media speculate that they will possibly ***reach*** (**PS+**) an agreement.
…	
Document-level Event Factuality: CT-	

Figure 1: An example document with both sentence- and document-level event factuality. The factuality between sentence- and document-level may be different.

Existing EFI studies mainly focus on sentence-level EFI, i.e., judging event factuality based on an individual sentence in which the event is located. In recent years, various neural models have been proposed for sentence-level EFI, and achieve state-of-the-art performance (Rudinger et al., 2018; Qian et al., 2018; Veyseh et al., 2019). Despite these successful efforts, sentence-level EFI suffers from an inevitable restriction in practice: it easily leads to conflicts between different mentions of the same event. Take Figure 1 as an example, the "*reach*" event is mentioned multiple times in a document, which has various factuality values in different sentences. The factuality of the event "*reach*" in S2 is PS+ according to the speculative word "*may*",

*Equal contribution.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2636–2645
November 7–11, 2021.

while in S3, its factuality is CT- due to the negative word "*denied*". According to our statistics on the English and Chinese event factuality datasets (Qian et al., 2019), 25.7% (English) and 37.8% (Chinese) of instances have the problem of event factuality conflict at sentence level for the same event, which is not negligible. Fortunately, the event factuality can be uniquely determined from the perspective of a document, which is able to naturally address the problem of sentence-level event factuality inconsistency. Therefore, it is necessary to move EFI forward from sentence level to document level.

However, identifying document-level event factuality is non-trivial. As shown in Figure 1, the factuality between document-level and sentence-level may be quite different. In this scenario, document-level event factuality cannot be deduced from each sentence-level factuality separately, but depends on the comprehensive semantic information of the entire document. To this end, we first learn the local information, and then integrate local representations to the global representation for prediction. In this process, we need to consider two important characteristics: *Local Uncertainty* and *Global Structure*, which can be leveraged to improve performance. In the following, we will introduce the two characteristics and give the reasons why they are critical for document-level EFI.

Local Uncertainty: As illustrated in Figure 1, different sentences (i.e., local information) describe different cognitive individuals' judgements towards the event factuality. However, the degree of uncertainty of these judgements is different. For example, as direct participants in the "*reach*" event, Mexican officials (in S4) can judge the event factuality with lower uncertainty (i.e., higher confidence) than other cognitive individuals (e.g., a journalist in S2). Apparently, the information of S4 is more important than that of S2 when predicting the document-level event factuality. It would be better if we could explicitly model the uncertainty of local information. Therefore, the first challenging problem is how to model the uncertainty of local information.

Global Structure: When integrating local information, utilizing global structure (i.e., document structure) could yield a better global representation. The global structure is manifested in two aspects: positional structure and semantic structure. For positional structure, as shown in Figure 1, the content of the document is roughly organized in chronological order, which can reflect the evolution of events. For semantic structure, there is a semantic correlation between local information. For instance, the content of S2 is the support of the view about the event factuality in S1, while the content of S3 is the denial of that in S1. There is no doubt that capturing the global structure enables a better understanding of documents. Thus, the second challenging problem is how to leverage the document structure for integrating local information.

In this paper, we propose a novel method termed as **U**ncertain **L**ocal-to-**G**lobal **N**etwork (ULGN) to address aforementioned problems. Specifically, to model the uncertainty of local information, we propose a *Local Uncertainty Estimation* module. It utilizes a probability distribution to represent the local information, rather than a deterministic feature vector. For ease of modeling, we adopt Gaussian distributions. Namely, the local information is now parameterized by a mean and variance. The former acts like the normal feature vector as in the conventional model, whereas the latter measures the feature uncertainty. The higher the uncertainty of the local information is, the larger its corresponding variance is. To leverage the global structure for synthesizing local information, we devise an *Uncertain Information Aggregation* module. The module first constructs a global graph based on the document structure, and then employs an uncertain graph convolution layer to aggregate the local information. It considers the uncertainty of local information via variance-based attention. Experimental results on two widely used datasets demonstrate that our method substantially outperforms previous state-of-the-art models.

Overall, the main contributions of this work can be summarized as follows:

- We propose a novel Uncertain Local-to-Global Network (ULGN) for document-level event factuality identification. To our best knowledge, we are the first to consider local uncertainty and global structure for the task.
- To model the uncertainty of local information, we propose a local uncertainty estimation module. To leverage the global structure for integrating local information, we devise an uncertain information aggregation module.
- Experimental results indicate that our approach significantly outperforms previous state-of-the-art methods, achieving 8.4% and

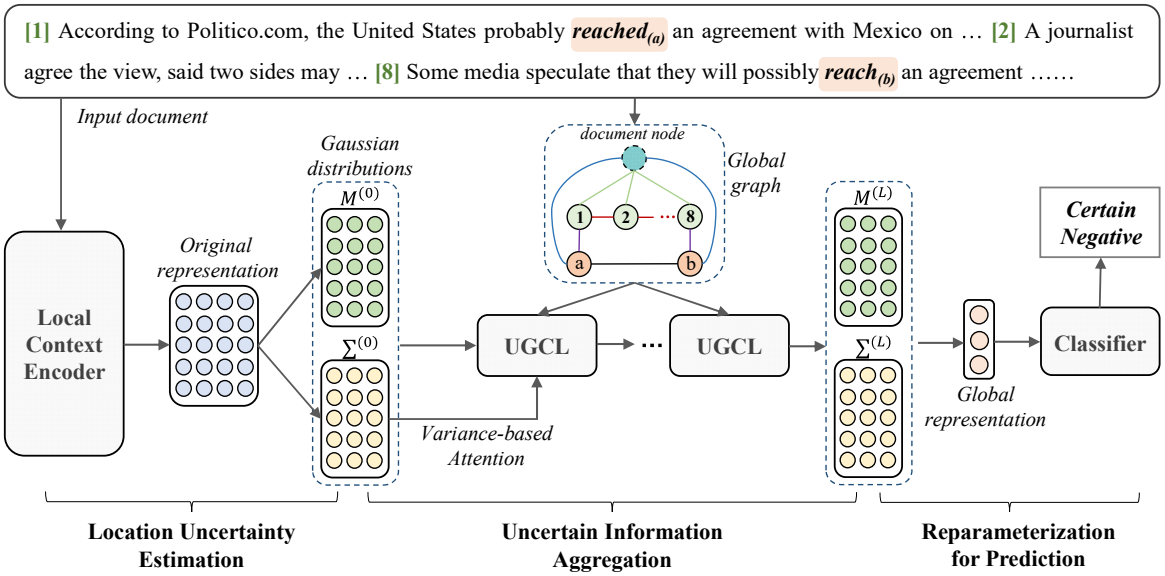

Figure 2: The architecture of our proposed uncertain local-to-global network for document-level event factuality identification. UGCL represents Uncertain Graph Convolution Layer introduced in Section 2.2.2.

11.45% improvements of F1 score on two widely used datasets. The source code of this paper is available at https://github.com/CPF-NLPR/ULGN4DocEFI.

2 Methodology

We propose an uncertain local-to-global network (ULGN) for document-level EFI. Figure 2 schematically visualizes our approach, which consists of three major components: (1) *Local Uncertainty Estimation* (§2.1), which represents the local information by using a probability distribution; (2) *Uncertain Information Aggregation* (§2.2), which leverages the global structure to integrate the local information; (3) *Reparameterization for Prediction* (§2.3), which utilizes the reparameterization trick (Kingma and Welling, 2013) to obtain the global representation for final prediction. We will illustrate each component in detail.

2.1 Local Uncertainty Estimation

We treat sentences and event mentions as the local information. Our local context encoder is based on the Transformer architecture (Vaswani et al., 2017). We adopt the BERT (Devlin et al., 2019) to encode the local information,[1] which has achieved the state-of-the-art performance for EFI task (Veyseh et al., 2019). The local context encoder takes each sentence of a document as input, which is defined as follows:

$$\boldsymbol{f}_i^s = \text{BERT}(S_i), \quad i = 1, 2, \ldots, N_s \tag{1}$$

where S_i denotes the i-th sentence and N_s is the number of sentences in the document. We use the [CLS] token representation of the last layer in BERT as the sentence representation. The representation of the event mention e_i ($i = 1, 2, \ldots, N_e$, where N_e is the number of times the event is mentioned.) is defined by averaging the representations of contained words, denoted as $\boldsymbol{f}_i^e$.

After obtaining the feature vector of the local information, we need to estimate its uncertainty. To this end, we use a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2)$ to represent the local information, instead of a deterministic feature vector. The $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i^2$ refer to mean vector and variance matrix respectively, which is formulated as follows:

$$\boldsymbol{\mu}_i = \boldsymbol{W}_\mu \boldsymbol{f}_i, \quad \boldsymbol{\sigma}_i^2 = \boldsymbol{W}_\sigma \boldsymbol{f}_i, \tag{2}$$

where $\boldsymbol{f}_i$ denotes the original representation of the sentence or event mention (i.e., $\boldsymbol{f}_i^s$ or $\boldsymbol{f}_i^e$). $\boldsymbol{W}_\mu$ and $\boldsymbol{W}_\sigma$ are trainable parameters.

In this way, each local information is represented as $\boldsymbol{h}_i = \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2)$ ($i = 1, 2, \ldots, N_s + N_e$), which not only gives the contextualized representation of local information (i.e., mean vector), but also estimates the uncertainty of local information (i.e., variance matrix).

[1] Note that the encoder is not our focus in this paper. In fact, other models like convolutional neural networks (Zeng et al., 2014) and long short-term memory networks (Hochreiter and Schmidhuber, 1997) can also be employed as encoders.

2.2 Uncertain Information Aggregation

2.2.1 Global Graph Construction

To leverage the global structure for integrating the local information, we first construct a *Global Graph*. As shown in Figure 2, the graph has three kinds of nodes: *mention node*, *sentence node* and *document node*. The mention nodes and sentence nodes can provide the local information for prediction. The global graph has one document node that aims to capture the information of the entire document. According to the document structure, we define the following five types of edges:

- *Adjacent sentence edge*: We connect a sentence node with its previous and next sentence nodes.
- *Document-sentence edge*: We connect the document node with all sentence nodes.
- *Document-mention edge*: All event mention nodes are connected to the document node.
- *Sentence-mention edge*: The mention node is connected to its corresponding sentence node.
- *Mention coreference edge*: Mentions referring to the same event are fully connected.

With the above connections, the positional structure can be modeled via the adjacent sentence edge. Besides, the document node could serve as a pivot to interact with other nodes and thus reduce the long distance among them in the document. Any two local nodes (i.e., mention nodes and sentence nodes) that are not directly connected can pass information to each other through the document node. Thus, the above connections can also model the semantic structure.

2.2.2 Uncertain Graph Convolution Layer

After constructing the global graph, we aggregate the local information based on the graph. For conventional graph convolution networks (GCNs) (Kipf and Welling, 2017), the (l+1)-th convolution layer is defined as:

$$\boldsymbol{h}_i^{(l+1)} = \rho\Big(\sum_{j\in ne(i)} \frac{1}{\sqrt{\tilde{\boldsymbol{D}}_{i,i}\tilde{\boldsymbol{D}}_{j,j}}} \boldsymbol{h}_j^{(l)} \boldsymbol{W}^{(l)}\Big), \tag{3}$$

or in the equivalent matrix form:

$$\boldsymbol{H}^{(l+1)} = \rho(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}), \tag{4}$$

where $\tilde{\boldsymbol{A}} = \boldsymbol{A}+\boldsymbol{I}$. $\boldsymbol{A}$ denotes the adjacency matrix of the global graph, and $\boldsymbol{I}$ is the identify matrix. $\tilde{\boldsymbol{D}}_{i,i} = \sum_j \tilde{\boldsymbol{A}}_{i,j}$. ρ is an activation function (e.g., ReLU). $ne(i)$ denotes neighbors of the node i.

Since the local information is parameterized by a probability distribution, existing graph convolutions are no longer applicable. Inspired by Zhu et al. (2019), we formally utilize an uncertain graph convolution layer (UGCL) to perform convolution operations between Gaussian distributions. Denote $\boldsymbol{h}_i^{(l)} = \mathcal{N}(\boldsymbol{\mu}_i^{(l)}, \boldsymbol{\sigma}_i^{(l)})$ as the representation of node i in l-th layer, where $\boldsymbol{\mu}_i^{(l)}$ is the mean vector and $\boldsymbol{\sigma}_i^{(l)}$ is the diagonal variance matrix[2]. We use $\boldsymbol{M}^{(l)} = [\boldsymbol{\mu}_1^{(l)}, \ldots, \boldsymbol{\mu}_{N_n}^{(l)}]$ and $\boldsymbol{\Sigma}^{(l)} = [\boldsymbol{\sigma}_1^{(l)}, \ldots, \boldsymbol{\sigma}_{N_n}^{(l)}]$ to denote the matrix of means and variances for all nodes respectively, where N_n is the number of nodes in the global graph (i.e., $N_n = N_s + N_e + 1$).

According to the additivity of the Gaussian distribution (LeCam, 1965) and assuming all hidden representations of nodes are independent, we can aggregate node neighbors as follows:

$$\begin{aligned}\boldsymbol{h}_{ne(i)}^{(l)} &= \sum_{j\in ne(i)} \frac{1}{\sqrt{\tilde{\boldsymbol{D}}_{i,i}\tilde{\boldsymbol{D}}_{j,j}}} \boldsymbol{h}_j^{(l)} \\ &\sim \mathcal{N}\Big(\sum_{j\in ne(i)} \frac{1}{\sqrt{\tilde{\boldsymbol{D}}_{i,i}\tilde{\boldsymbol{D}}_{j,j}}} \boldsymbol{\mu}_j^{(l)}, \sum_{j\in ne(i)} \frac{1}{\tilde{\boldsymbol{D}}_{i,i}\tilde{\boldsymbol{D}}_{j,j}} \boldsymbol{\sigma}_j^{(l)}\Big).\end{aligned} \tag{5}$$

Due to the different importance of the local information, we propose a variance-based attention mechanism to assign different weights to neighbors. Intuitively, a smaller variance means that the node is more important. Specifically, we use a smooth exponential function to control the effect of variances on weight:

$$\boldsymbol{\alpha}_i^{(l)} = \exp(-\gamma \boldsymbol{\sigma}_i^{(l)}), \tag{6}$$

where $\boldsymbol{\alpha}_i^{(l)}$ are the attention weights of node i in the l-th layer and γ is a hyper-parameter. Considering the variance-based attention, the Eq.(5) can be modified as follows:

$$\begin{aligned}\boldsymbol{h}_{ne(i)}^{(l)} &= \sum_{j\in ne(i)} \frac{1}{\sqrt{\tilde{\boldsymbol{D}}_{i,i}\tilde{\boldsymbol{D}}_{j,j}}} (\boldsymbol{h}_j^{(l)} \odot \boldsymbol{\alpha}_j^{(l)}) \sim \\ &\mathcal{N}\Big(\sum_{j\in ne(i)} \frac{\boldsymbol{\mu}_j^{(l)} \odot \boldsymbol{\alpha}_j^{(l)}}{\sqrt{\tilde{\boldsymbol{D}}_{i,i}\tilde{\boldsymbol{D}}_{j,j}}}, \sum_{j\in ne(i)} \frac{\boldsymbol{\sigma}_j^{(l)} \odot \boldsymbol{\alpha}_j^{(l)} \odot \boldsymbol{\alpha}_j^{(l)}}{\tilde{\boldsymbol{D}}_{i,i}\tilde{\boldsymbol{D}}_{j,j}}\Big),\end{aligned} \tag{7}$$

where $\odot$ denotes the element-wise product operation. To better integrating the local information,

[2]In this paper, we focus on diagonal variance matrices, but according to Hoeffding (1994), it can be extended to more general cases. In addition, for the ease of presentation, we use $\boldsymbol{\sigma}$ to represent variances, instead of $\boldsymbol{\sigma}^2$.

the attention weights are exerted for different dimensions separately.

Similar to Eq.(3), we need to apply learnable filters and non-linear activation functions to $\boldsymbol{h}_{ne(i)}^{(l)}$ for obtaining $\boldsymbol{h}_i^{(l+1)}$. However, since $\boldsymbol{h}_{ne(i)}^{(l)}$ is a Gaussian distribution, it is mathematically intractable to compute $\boldsymbol{h}_i^{(l+1)}$. In such a scenario, we directly impose layer-specific parameters and non-linear activation functions to the means and variances, respectively. Therefore, the uncertain graph convolution can be defined as follows:

$$\begin{aligned}\boldsymbol{\mu}_i^{(l+1)} &= \rho\Big(\sum_{j\in ne(i)} \frac{1}{\sqrt{\tilde{\boldsymbol{D}}_{i,i}\tilde{\boldsymbol{D}}_{j,j}}}(\boldsymbol{\mu}_j^{(l)} \odot \boldsymbol{\alpha}_j^{(l)})\boldsymbol{W}_\mu^{(l)}\Big) \\ \boldsymbol{\sigma}_i^{(l+1)} &= \rho\Big(\sum_{j\in ne(i)} \frac{1}{\tilde{\boldsymbol{D}}_{i,i}\tilde{\boldsymbol{D}}_{j,j}}(\boldsymbol{\sigma}_j^{(l)} \odot \boldsymbol{\alpha}_j^{(l)} \odot \boldsymbol{\alpha}_j^{(l)})\boldsymbol{W}_\sigma^{(l)}\Big),\end{aligned} \tag{8}$$

or equivalently in the matrix form:

$$\begin{aligned}\boldsymbol{M}^{(l+1)} &= \rho(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}(\boldsymbol{H}^{(l)} \odot \mathcal{A}^{(l)})\boldsymbol{W}_\mu^{(l)}) \\ \boldsymbol{\Sigma}^{(l+1)} &= \rho(\tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-1}(\boldsymbol{\Sigma}^{(l)} \odot \mathcal{A}^{(l)} \odot \mathcal{A}^{(l)})\boldsymbol{W}_\sigma^{(l)}),\end{aligned} \tag{9}$$

where $\mathcal{A}^{(l)} = \exp(-\gamma\boldsymbol{\Sigma}^{(l)})$. $\boldsymbol{M}^{(0)}$ and $\boldsymbol{\Sigma}^{(0)}$ are computed via Eq.(2).

2.3 Reparameterization for Prediction

We use the representation of the document node for prediction. Considering the representation of the document node is a Gaussian distribution, we first adopt a sampling process in the last graph layer:

$$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{u}_d^{(L)}, \boldsymbol{\sigma}_d^{(L)}), \tag{10}$$

where $\mathcal{N}(\boldsymbol{u}_d^{(L)}, \boldsymbol{\sigma}_d^{(L)})$ denotes the representation of the document node in the last layer. However, directly sampling $\boldsymbol{z}$ will cause the problem of preventing gradients from propagating back to the preceding layers. Thus, we use the reparameterization trick (Kingma and Welling, 2013) to bypass the problem. Specifically, we first sample a random noise $\boldsymbol{\epsilon}$ from the standard Gaussian distribution, and then generate $\boldsymbol{z}$ as the equivalent sampling representation:

$$\boldsymbol{z} = \boldsymbol{u}_d^{(L)} + \boldsymbol{\epsilon} \odot \sqrt{\boldsymbol{\sigma}_d^{(L)}}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}). \tag{11}$$

After obtaining $\boldsymbol{z}$, we feed it into a softmax function for prediction:

$$\boldsymbol{p} = \text{softmax}(\boldsymbol{W}_s\boldsymbol{z} + \boldsymbol{b}_s). \tag{12}$$

For training, we use the cross entropy loss to optimize the model parameters:

$$\mathcal{L}_{cls} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{y}_i \cdot \log(\boldsymbol{p}), \tag{13}$$

where N is the number of training instances. $\boldsymbol{y}_i$ is the label of the i-th instance.

In addition, to ensure that the learned representations are indeed Gaussian distributions, we devise an explicit regularization loss to constrain the input representations of the first layer:

$$\mathcal{L}_{reg} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N_n} \text{KL}(\mathcal{N}(\boldsymbol{u}_{ij}^{(0)}, \boldsymbol{\sigma}_{ij}^{(0)})||\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})), \tag{14}$$

where $\mathcal{N}(\boldsymbol{u}_{ij}^{(0)}, \boldsymbol{\sigma}_{ij}^{(0)})$ is the initialized Gaussian distribution of j-th node of i-th instance. $\text{KL}(\cdot||\cdot)$ is the KL-divergence between two distribution. Since deeper layers are naturally Gaussian distributions by using the proposed UGCL, we only need to regularize $\boldsymbol{M}^{(0)}$ and $\boldsymbol{\Sigma}^{(0)}$. We reach the final loss function by combining the above terms:

$$\mathcal{L} = \mathcal{L}_{cls} + \beta\mathcal{L}_{reg}, \tag{15}$$

where β is a hyper-parameter that controls the impact of the $\mathcal{L}_{reg}$.

3 Experiments

3.1 Datasets and Evaluation Metrics

We evaluate our proposed method on two widely used datasets, English and Chinese event factuality datasets (Qian et al., 2019). The number of English and Chinese documents is 1,730 and 4,650, respectively. The PS- and Uu documents only cover 1.39% and 1.20% in the English and Chinese datasets, respectively. Therefore, following previous work (Qian et al., 2019), we mainly focus on the performance of CT+, CT- and PS+. For a fair comparison with previous work (Qian et al., 2019), we both perform 10-fold cross-validation on English and Chinese corpora. In addition, we adopt F1 score as the evaluation metric for each category of factuality value. We also consider macro-averaged and micro-averaged F1 score for the overall performance of all the categories of factuality values.

3.2 Parameter Settings

In our implementations, our method uses the HuggingFace's Transformers library[3] to implement the BERT base model, which has 12-layers, 768-hidden, and 12-heads. The learning rate is initialized as 2e-5 with a linear decay. We use the AdamW algorithm (Loshchilov and Hutter, 2018) to optimize model parameters. The batch size is set

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

Datasets	Methods	CT+ (%)	CT- (%)	PS+ (%)	Micro-F1 (%)	Macro-F1 (%)
English	BERT Model	89.38	71.82	69.09	83.53	76.76
	MaxEntVote	75.14	58.17	35.89	68.42	56.40
	BiLSTM-Att	79.18	65.25	53.65	73.23	66.03
	Att-Adv	89.84	76.87	62.14	83.56	76.28
	ULGN (Ours)	**92.49** (↑ 2.65)	**84.87** (↑ 8.00)	**76.68** (↑ 14.54)	**88.69** (↑ 5.13)	**84.68** (↑ 8.40)
Chinese	BERT Model	84.79	88.71	79.33	85.83	84.28
	MaxEntVote	72.22	62.44	58.29	67.72	64.32
	BiLSTM-Att	81.89	68.82	49.78	71.12	67.28
	Att-Adv	87.52	83.35	74.06	84.03	81.64
	ULGN (Ours)	**93.53** (↑ 6.01)	**94.99** (↑ 11.64)	**90.76** (↑ 16.70)	**93.77** (↑ 9.74)	**93.09** (↑ 11.45)

Table 1: Experimental results on the English and Chinese event factuality datasets, respectively. The performance of our method is followed by the improvements (↑) over the previous state-of-the-art method Att-Adv.

to 4 and 2 for English and Chinese event factuality datasets, respectively. The number of uncertain graph convolution layers is set to 2. The size of hidden states of the uncertain graph convolution layer is 768.

3.3 Baselines

We compare the proposed approach ULGN with the following methods:

(1) **MaxEntVote** (Qian et al., 2019), which first uses maximum entropy model to identify sentence-level event factuality, and then votes, i.e., choosing the value committed by the most sentences as the document-level factuality value.

(2) **BiLSTM-Att** (Qian et al., 2019), which employs the bidirectional long short-term memory network (BiLSTM) to extract features, and uses the intra-sentence attention to capture the most important information in the sentence.

(3) **Att-Adv** (Qian et al., 2019), which leverages the intra-sentence and inter-sentence attention to learn the document representation, and utilizes adversarial training to improve the robustness.

(4) **BERT Model**, which utilizes the BERT-base (Devlin et al., 2019) to encode the document, and uses the [CLS] representation for prediction.

3.4 Overall Results

Table 1 shows the results on the English and Chinese datasets, respectively. We note the following key observations throughout our experiments:

(1) Our method outperforms all the baselines by a large margin. For example, compared with the previous state-of-the-art model Att-Adv (Qian et al., 2019), our method achieves 11.45% improvements of macro-F1 score on the Chinese event factuality dataset. The significant performance gain of our method over the baselines demonstrates that the proposed ULGN is very effective for this task.

(2) Our method improves upon the BERT Model by 7.92% and 8.81% in term of macro-F1 score on the English and Chinese event factuality datasets, respectively. We attribute the improvements to that our method ULGN takes advantage of local uncertainty and global structure, thus achieving superior performance than the BERT Model.

(3) The BERT Model achieves comparable performance with complex state-of-the-art methods such as Att-Adv (Qian et al., 2019) on these two datasets, which indicates that the BERT is able to extract useful text features for the task.

3.5 Ablation Study

To demonstrate the effectiveness of the local uncertainty estimation (LUE) and uncertain information aggregation (UIA), we conduct an ablation study as follows. 1) w/o VA, which removes the variance-based attention; 2) w/o LUE, which first uses BERT to encode the local information as the vector, and then employs vanilla GCNs to aggregate the local information; 3) w/o UIA, which first samples a representation (i.e., vector) for each local information, and then performs max-pooling over these sampled representations to get the global representation for prediction; 4) w/o LUE and UIA, which is the same as the BERT Model introduced in Section 3.3. We present the results of ablation study in Table 2. From the results, we can observe that:

(1) **Effectiveness of Local Uncertainty Estimation.** When we remove the LUE module from the ULGN, the macro-F1 score drops by 4.35% on

Datasets	Methods	CT+ (%)	CT- (%)	PS+ (%)	Micro-F1 (%)	Macro-F1 (%)
English	ULGN	**92.49**	**84.87**	**76.68**	**88.69**	**84.68**
	w/o VA	91.68 (↓ 0.81)	80.59 (↓ 4.28)	74.30 (↓ 2.38)	87.12 (↓ 1.57)	82.19 (↓ 2.49)
	w/o LUE	89.45 (↓ 3.04)	81.44 (↓ 3.43)	70.11 (↓ 6.57)	85.18 (↓ 3.51)	80.33 (↓ 4.35)
	w/o UIA	89.31 (↓ 3.18)	79.86 (↓ 5.01)	69.32 (↓ 7.36)	84.74 (↓ 3.95)	79.50 (↓ 5.18)
	w/o LUE and UIA	89.38 (↓ 3.11)	71.82 (↓ 13.05)	69.09 (↓ 7.59)	83.53 (↓ 5.16)	76.76 (↓ 7.92)
Chinese	ULGN	**93.53**	**94.99**	**90.76**	**93.77**	**93.09**
	w/o VA	92.62 (↓ 0.91)	94.30 (↓ 0.69)	87.41 (↓ 3.35)	92.53 (↓ 1.24)	91.44 (↓ 1.65)
	w/o LUE	89.34 (↓ 4.19)	92.47 (↓ 2.52)	86.56 (↓ 4.20)	90.45 (↓ 3.32)	89.46 (↓ 3.63)
	w/o UIA	88.06 (↓ 5.47)	91.22 (↓ 3.77)	85.12 (↓ 5.64)	88.23 (↓ 5.54)	88.13 (↓ 4.96)
	w/o LUE and UIA	84.79 (↓ 8.74)	88.71 (↓ 6.28)	79.33 (↓ 11.43)	85.83 (↓ 7.94)	84.28 (↓ 8.81)

Table 2: Ablation study by removing the main components, where "w/o" indicates without. The performance is followed by the drop (↓) compared with the method ULGN. "VA", "LUE" and "UIA" refer to "variance-based attention", "local uncertainty estimation" and "uncertain information aggregation", respectively.

n	Methods	Micro-F1 (%)	Macro-F1 (%)
n=1	Att-Adv	91.36	81.67
	ULGN	**92.48** (↑ 1.12)	**85.21** (↑ 3.54)
n>1	Att-Adv	60.91	60.04
	ULGN	**75.51** (↑ 14.60)	**74.76** (↑ 14.72)

Table 3: Experimental results of Att-Adv and our method ULGN on the documents with n types of sentence-level factuality values in the English dataset.

Methods	Micro-F1 (%)	Macro-F1 (%)
BERT Model	83.53	76.76
Longformer Model	83.81	77.48
BERT-GCN	85.18	80.33
BERT-GAT	85.22	80.41
ULGN (Ours)	**88.69** (↑ 3.47)	**84.68** (↑ 4.27)

Table 4: Comparison between the different methods for document modeling on the English dataset.

the English dataset. It proves the local uncertainty estimation is very effective for the task.

(2) **Effectiveness of Uncertain Information Aggregation.** Compared with the model removed UIA module, our method ULGN achieves 5.54% improvements of micro-F1 score on the Chinese dataset. Moreover, removing the VA module also brings performance degradation. It demonstrates that the uncertain information aggregation is able to effectively integrate the local information.

(3) **Effectiveness of Local Uncertainty Estimation and Uncertain Information Aggregation.** When we remove the LUE and UIA, the performance drops significantly. The macro-F1 score drops from 93.09% to 84.28% on the Chinese dataset. It indicates simultaneously utilizing the local uncertainty estimation and uncertain information aggregation is also very effective.

3.6 Results on the Documents with Different Sentence-Level Event Factuality Values

The document-level EFI is very challenging, because a document may have different sentence-level event factuality values. To further investigate the effectiveness of our method for document-level EFI, we compare our method with Att-Adv on the documents with n types of sentence-level factuality values. The results are shown in Table 3. From the table, we have two important observations:

(1) Compared with improvements over the Att-Adv (Qian et al., 2019) when n=1, our method achieves more improvements when n>1. For example, our method ULGN achieves 3.54% improvements of macro-F1 score when n=1, while 14.72% improvement when n>1 on the English dataset. It indicates that our method is able to handle well the problem of sentence-level event factuality inconsistency.

(2) The micro-F1 and macro-F1 of n>1 are lower than those of n=1 for both Att-Adv and our approach ULGN, indicating that the factuality of documents that have different types of sentence-level factuality are more difficult to identify due to the interference from sentence-level values.

3.7 Discussion and Analysis

3.7.1 Different Methods for Document Modeling

To validate the effectiveness of our method for document modeling, we compare our method with other baselines. The baselines are illustrated as follows. 1) Longformer Model, which uses the

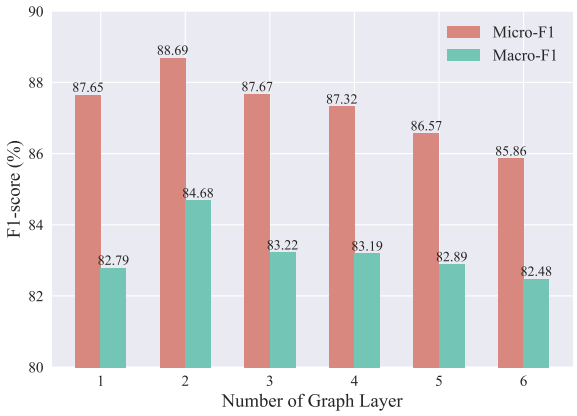

Figure 3: Influences of numbers of the graph layer on the English event factuality dataset.

Longformer[4] (Beltagy et al., 2020) to extract the global feature for prediction; 2) BERT-GCN and BERT-GAT, which first uses the BERT to the local information, and then employs GCN and GAT (Veličković et al., 2017) for integrating the local information, respectively.

We present the experimental results in Table 4. From the results, we can clearly see that our method ULGN significantly outperforms other baselines. It indicates that when modeling the document for the document-level EFI task, we not only need to consider the uncertainty of local information, but also need to leverage the document structure for integrating local information.

3.7.2 Impact of the Number of Graph Layers

We evaluate the influence of graph layer numbers, which is illustrated in Figure 3. From the figure, we can observe that:

(1) Our method ULGN yields the best performance when the number of graph layers is 2. We attribute it to the fact that any two local nodes that are not directly connected can pass information to each other through the document node (i.e., 2-hop).

(2) When the number of graph layers is too large, the micro-F1 and macro-F1 scores both stop increasing or even decrease. We guess that increasing the size of randomly initialized parameters may not be beneficial for BERT fine-tuning.

4 Related Work

4.1 Event Factuality Identification

Event factuality identification (EFI) is a very important task in information extraction, which can benefit many NLP applications, including rumor detection (Qazvinian et al., 2011), sentiment analysis (Klenner and Clematide, 2016), event causality identification (Cao et al., 2021; Tran Phu and Nguyen, 2021) and so on. Therefore, it has attracted extensive attention among researchers. Most existing EFI studies are limited to the sentence-level task (Saurí and Pustejovsky, 2012; De Marneffe et al., 2012; Rudinger et al., 2018; Veyseh et al., 2019). The early work on this problem has mainly employed rule-based methods (Nairn et al., 2006; Saurı, 2008; Lotan et al., 2013) or machine learning methods (with manually designed features) (Diab et al., 2009; Prabhakaran et al., 2010; De Marneffe et al., 2012; Saurí and Pustejovsky, 2012; Lee et al., 2015; Qian et al., 2015). In recent years, neural networks have been introduced into the EFI task, and achieved state-of-the-art performance (Rudinger et al., 2018; Qian et al., 2018; Sheng et al., 2019; Huang et al., 2019; Veyseh et al., 2019).

Despite these successful efforts, sentence-level event factuality easily leads to conflict. To this end, Qian et al. (2019) propose the document-level EFI task. However, when modeling the document for the task, their method ignores the uncertainty of local information and the global structure.

4.2 Uncertainty Modeling

Uncertainty is a crucial but long-ignored issue in many applications of NLP area. Conventionally, the high-level representation of an input instance is modeled as a fixed-length feature vector, which can be regarded as a "*point*" in low-dimensional spaces. However, such a point estimate is not sufficient to express uncertainty, as point-based methods assume that learned features are always correct (Gal and Ghahramani, 2016; Kendall and Gal, 2017). In recent years, Gaussian embedding has been getting more attention in deep learning. For example, Vilnis and McCallum (2015) utilize Gaussian embeddings to represent words, where the covariance naturally measures the ambiguity of the words. He et al. (2015) attempt to leverage the Gaussian distribution to represent the entity and relation, which aims to model the uncertainty of entities and relations in knowledge graphs. In addition, Xiao and Wang (2019) quantify uncertainties in some NLP tasks, such as sentiment analysis, named entity recognition and language modeling.

To the best of our knowledge, we are the first to

[4]Longformer can model longer texts than BERT. The maximum length it can handle is 4,096.

consider the uncertainty of local information for the document-level EFI task. Namely, we represent the local information as a probability distribution, rather than a deterministic feature vector.

5 Conclusion

In this paper, we propose a novel uncertain local-to-global network (ULGN) for document-level event factuality identification. To model the uncertainty of local information, we propose a local uncertainty estimation module to represent the local information with a probability distribution. To leverage the global structure, we devise an uncertain information aggregation module to integrate the local information. Experimental results on two widely used datasets indicate that our approach substantially outperforms previous state-of-the-art methods.

Acknowledgments

We thank anonymous reviewers for their insightful comments and suggestions. This work is supported by the National Natural Science Foundation of China (No.U1936207, No.61976211, No.61806201), the Beijing Academy of Artificial Intelligence (BAAI2019QN0301), the Key Research Program of the Chinese Academy of Sciences (Grant NO. ZDBS-SSW-JSC006), the independent research project of National Laboratory of Pattern Recognition and the Youth Innovation Promotion Association CAS.

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, Yuguang Chen, and Weihua Peng. 2021. Knowledge-enriched event causality identification via latent structure induction networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4862–4872. Association for Computational Linguistics.

Marie-Catherine De Marneffe, Christopher D Manning, and Christopher Potts. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Computational 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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Mona Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 68–73.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of International Conference on Machine Learning*, pages 1050–1059. PMLR.

Shizhu He, Kang Liu, Guoliang Ji, and Jun Zhao. 2015. Learning to represent knowledge graphs with gaussian embedding. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 623–632.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Wassily Hoeffding. 1994. On sequences of sums of independent random vectors. In *The Collected Works of Wassily Hoeffding*, pages 395–408. Springer.

Rongtao Huang, Bowei Zou, Hongling Wang, Peifeng Li, and Guodong Zhou. 2019. Event factuality detection in discourse. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 404–414. Springer.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5580–5590.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, 2017,Conference Track Proceedings*.

Manfred Klenner and Simon Clematide. 2016. How factuality determines sentiment inferences. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 75–84.

Lucien LeCam. 1965. On the distribution of sums of independent random variables. In *Bernoulli 1713, Bayes 1763, Laplace 1813*, pages 179–202. Springer.

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1648.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Amnon Lotan, Asher Stern, and Ido Dagan. 2013. Truthteller: Annotating predicate truth. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 752–757.

Rowan Nairn, Cleo Condoravdi, and Lauri Karttunen. 2006. Computing relative polarity for textual inference. In *Proceedings of the fifth international workshop on inference in computational semantics*.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2010. Automatic committed belief tagging. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1014–1022.

Vahed Qazvinian, Emily Rosengren, Dragomir Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599.

Zhong Qian, Peifeng Li, Yue Zhang, Guodong Zhou, and Qiaoming Zhu. 2018. Event factuality identification via generative adversarial networks with auxiliary classification. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 4293–4300.

Zhong Qian, Peifeng Li, and Qiaoming Zhu. 2015. A two-step approach for event factuality identification. In *2015 International Conference on Asian Language Processing*, pages 103–106. IEEE.

Zhong Qian, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2019. Document-level event factuality identification via adversarial neural network. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2799–2809.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 193–203.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural models of factuality. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 731–744.

Roser Saurı. 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. *Computational linguistics*, 38(2):261–299.

Jiaxuan Sheng, Bowei Zou, Zhengxian Gong, Yu Hong, and Guodong Zhou. 2019. Chinese event factuality detection. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 486–496. Springer.

Minh Tran Phu and Thien Huu Nguyen. 2021. Graph convolutional networks for event causality identification with rich document-level structures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3480–3490. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. In *Proceedings of International Conference on Learning Representations*.

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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4393–4399.

Luke Vilnis and Andrew McCallum. 2015. Word representations via gaussian embedding. In *Proceedings of International Conference on Learning Representations*.

Yijun Xiao and William Yang Wang. 2019. Quantifying uncertainties in natural language processing tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7322–7329.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of the 25th international conference on computational linguistics: technical papers*, pages 2335–2344.

Dingyuan Zhu, Ziwei Zhang, Peng Cui, and Wenwu Zhu. 2019. Robust graph convolutional networks against adversarial attacks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1399–1407.