ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230508

Uncertain Relational Hypergraph Attention Networks for Document-Level Event Factuality Identification

Jiawei Sheng[1,2], Xin Cong[1,2], Jiangxia Cao[1,2], Shu Guo[3], Chen Li[4], Lihong Wang[3], Tingwen Liu[1,2], and Hongbo Xu[1,2]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
[3]CNCERT / CC [4]ARC Lab, Tencent
shengjiawei@iie.ac.cn, guoshu@cert.org.cn, wlh@isc.org.cn, liutingwen@iie.ac.cn, hbxu@iie.ac.cn

Abstract. Document-level event factuality identification (DocEFI) is an important task in event knowledge acquisition, which aims to detect whether an event actually occurs or not from the perspective of the document. Unlike the sentence-level task, a document can have multiple sentences with different event factualities, leading to event factuality conflicts in DocEFI. Existing studies attempt to aggregate local event factuality by exploiting document structures, but they mostly consider textual components in the document separately, degrading complicated correlations therein. To address the above issues, this paper proposes a novel approach, namely URHAT, to improve DocEFI with uncertain relational hypergraph attention networks. Particularly, we reframe a document graph as a hypergraph, and establish beneficial n-ary correlations among textual nodes with relational hyperedges, which helps to globally consider local factuality features to resolve event factuality conflicts. To better discern the importance of event factuality features, we further represent textual nodes with uncertain Gaussian distributions, and propose novel uncertain relational hypergraph attention networks to refine textual nodes with the document hypergraph. In addition, we select factuality-related keywords as nodes to enrich event factuality features. Experimental results demonstrate the effectiveness of our proposed method, and outperforms previous methods on two widely used benchmark datasets.

Document-level Event: *reach_agreement*

Sentences:

(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 of a document-level event with category CT-. The mentions in sentences can have different categories with the document.

1 Introduction

Document-level event factuality identification (DocEFI) aims to identify the event factuality of a target event in the document. Generally, **event factuality** refers to the degree of certainty about whether an event actually happens or not from the perspective of a given text, which is usually classified into five categories [39]: certainly happens (CT+), certainly does not happen (CT-), possibly happens (PS+), possibly does not happen (PS-), and underspecified (Uu). As shown in Figure 1, the **event mention**[1] *reaches* in sentence S1 implies that the mentioned event *reach_agreement* might happen, but it is not sure, so the mention should be classified in the category PS+. Nevertheless, by considering all the mentions in their associated sentences, the model should output CT- as the document-level event factuality. In the context of the rapid development of large generative language models [27], such a task can help to acquire reliable event knowledge from massive generated contents, and facilitate several downstream applications, such as rumour detection [31], sentiment analysis [18], and event causality identification [5, 7].

Traditional studies [26, 33, 34, 29] mostly focus on the sentence-level task, which detects event factuality for sentences independently, and would suffer from the **event factuality conflicts** in DocEFI. As shown in Figure 1, the document contains several sentences with conflicting event factuality categories. S1, S2 and S8 associate the PS+ category, where the narrators express a positive standpoint and spec-

[1] The term is usually called event trigger in event extraction studies [6, 24, 43, 44]. Here we follow previous studies [39, 35, 4] in this task.

ulative attitude towards the event, considering the word *probably* and *may* as cues. However, S3 and S4 associate the CT category, which expresses the negative standpoint and clear attitude towards the event given by the official narrators. According to the statistics, there are 25.7% (English) and 37.8% (Chinese) of documents having event factuality conflicts in the DocEFI datasets [35]. The DocEFI task requires a global consideration of all the mentions in the sentences involved and factuality predictions from a document-level perspective. Such conflicting event factuality categories may confuse the model, making the task more challenging in real-world applications.

Existing DocEFI studies [35, 4] attempt to resolve the event factuality conflicts by capturing event correlations across sentences. Among them, Qian et al. [35] propose the DocEFI task, which adopts an attention mechanism to adaptively learn event correlations with implicit document structures. Then, Cao et al. [4] explicitly treat each document as a homogeneous graph, and formulate event correlations by establishing separate edges between the document and mention nodes. They treat node features as uncertain Gaussian distributions (*a.k.a.*, Gaussian embeddings) [47], and then employ uncertain graph convolutional networks [54] to differentiate factuality features. However, they consider the mentions referring to the same event separately, which degrades the coreference relationship among the mentions and thus hinders the understanding of the event factuality at the document-level.

We argue that it is crucial to fully exploit the complicated document structures for factuality prediction. First, a document consists of multi-grained textual components, including sentences, mentions and words, and a sentence may contain some keywords that explicitly imply the factuality of the event, such as the word *probably* for PS+ in Figure 1. There are also **n-ary correlations** among textual components (n is the number of textual components). For example, in Figure 1, all mentions refer to the same document-level event, where each mention reflects only a partial perspective on the factuality. Considering these mentions separately may degrade the entire n-ary correlation, namely *mentions-of-document*, and easily mislead the model about part of the mentions. Intuitively, the document requires to be globally correlated with local textual components in order to capture a global factuality perspective. Moreover, there are also correlations among local textual components that can refine local perspectives with other perspectives. Such complicated correlations can hardly be established by simple graph structures.

To address the above issues, we propose a novel approach for DocEFI with Uncertain Relational Hypergraph Attention networks, termed as UR-HAT. Specifically, to construct the complicated document structures, we formulated each document as a **hypergraph** [3], which contains hyperedges that can simultaneously connect multiple textual nodes, thus establishing n-ary correlations among textual components. In addition, we devise factuality-related keyword nodes to explicitly enrich factuality features. In order to discern the importance of nodes in correlations, we further represent node features as Gaussian distributions, and propose uncertain hypergraph networks to refine node features with the document hypergraph structure. In this way, the model not only retains the factuality features in the means, but also estimates the uncertainty of the factuality features in the variances, which helps to discern beneficial node features in n-ary correlations. In summary, the contributions are threefold:

- We propose a novel approach, namely UR-HAT, for the DocEFI task. To the best of our knowledge, we are the first to exploit hypergraphs to improve the understanding of event factuality at the document level compared to previous works.
- We technically propose an uncertain relational hypergraph attention network to aggregate event factuality features, which pioneeringly models the uncertainty of features in hypergraph learning.
- Experimental results indicate the effectiveness of our method[2], outperforming previous methods on two widely used datasets.

2 Preliminaries

Before going on, we briefly retrospect the hypergraph structure and paradigmatic hypergraph neural networks.

2.1 Hypergraph Structure

Unlike conventional graphs, a hypergraph [3] is a special form of graph that contains hyperedges that can connect two or more nodes and is typically used to represent n-ary correlations [10]. In general, a hypergraph is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{P})$, which contains a node set $\mathcal{V} = \{v_1, ..., v_n\}$, a hyperedge set $\mathcal{E} = \{e_1, ..., e_m\}$, and an optional diagonal weight matrix $\boldsymbol{P} \in \mathbb{R}^{m\times m}$ representing the hyperedge weight. The hypergraph $\mathcal{G}$ can be denoted by an incidence matrix $\boldsymbol{H} \in \{0, 1\}^{n\times m}$ with entries defined as follows:

$$\boldsymbol{H}_{i,j} = \begin{cases} 1, & \text{if } v_i \in e_j \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where each hyperedge e_j connects all the involved nodes v_i, representing the correlations among these nodes. In this paper, we borrow the idea of the hypergraph to represent n-ary correlations in the document, providing expressiveness to establish complicated correlations.

2.2 Hypergraph Neural Networks

Hypergraph neural networks [10, 13] have been proven to be powerful for hypergraph learning. A typical study is HGCN [10], which develops graph convolutional network (GCN) [17] upon hypergraphs. Formally, the hypergraph convolution process can be written as:

$$\boldsymbol{X}^{l+1} = \delta\big(\mathrm{Norm}(\boldsymbol{H})\boldsymbol{P}(\mathrm{Norm}(\boldsymbol{H}^T)\boldsymbol{X}^l\boldsymbol{W}^l)\big), \tag{2}$$

where δ is an activation function, and Norm is the row-normalization function for numerical stability. Actually, Eq. (2) first aggregates node features for hyperedge representation, and then aggregates hyperedge features for node representation. Therefore, Ding et al. [9] summarize a general spatial-based aggregation process:

$$\begin{aligned} \boldsymbol{x}_i^{l+1} &= \mathrm{AGGR}^l_{\text{edge}}(\boldsymbol{x}_i^l, \{\boldsymbol{e}_j^l | \forall e_j \in \mathcal{E}_i\}), \\ \text{where } \boldsymbol{e}_j^l &= \mathrm{AGGR}^l_{\text{node}}(\{\boldsymbol{x}_k^l | \forall v_k \in e_j\}), \end{aligned} \tag{3}$$

where $\mathcal{E}_i$ denotes the set of hyperedges connected to node v_i, and $\boldsymbol{e}_j^l$ is the feature of hyperedge e_j at layer l. Here $\mathrm{AGGR}_{\text{node}}$ aggregates node features to generate hyperedge features, and $\mathrm{AGGR}_{\text{edge}}$ aggregates hyperedge features to generate node features. In general, the aggregation function can also be achieved by pooling or attention mechanism [9]. In this paper, we develop the above aggregation process from traditional certain hypergraph neural networks into an uncertain manner, which provides more capability to consider the uncertainty of node or hyperedge features.

[2] Our source code is released at https://github.com/JiaweiSheng/UR-HAT, and the supplemental **Appendix** is also available there.

3 Approach

The goal of DocEFI is to predict event factuality category for a target event, considering all the event mentions of sentences in the document. To this end, we construct the document structure as a hypergraph, and further propose uncertain hypergraph networks to refine the document-level event factuality features for prediction. Figure 3 outlines the overall framework of UR-HAT.

3.1 *Document Hypergraph Construction*

To establish document structures, we first derive keywords as additional features, and then construct the document hypergraph.

3.1.1 *Factuality-Related Keyword Selection*

To explicitly exploit sentence-level factuality features, we retrospect canonical feature selection methods, and derive factuality-related keywords. Here we employ mutual information (MI) [25], which mathematically measures how much information the presence/absence of a word t contributes to the sentence factuality category c. Formally:

$$\mathrm{I}(T;C) = \sum_{s_t \in \{1,0\}} \sum_{s_c \in \{1,0\}} P(s_t, s_c) \log \frac{P(s_t, s_c)}{P(s_t)P(s_c)} \tag{4}$$

where T is a random variable taking $s_t = 1$ when the sentence contains word t; and C is a random variable taking $s_c = 1$ when the sentence is in the category c. To implement the MI scores, we rewrite Eq. (4) using maximum likelihood estimation (MLE):

$$\begin{aligned} &\mathrm{I}(T=t;C=c) \\ &= \frac{N_{11}}{N} \log \frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N} \log \frac{NN_{01}}{N_{0.}N_{.1}} \\ &+ \frac{N_{10}}{N} \log \frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N} \log \frac{NN_{00}}{N_{0.}N_{.0}}, \end{aligned} \tag{5}$$

where the N with subscripts denote the number of sentences having the value of s_t and s_c. For example, N_{10} is the number of sentences containing the word t (*i.e.*, $s_t = 1$) and not belonging to the category c (*i.e.*, $s_c = 0$). $N_{1.} = N_{10} + N_{11}$ is the number of all the sentences containing the word t (*i.e.*, $s_t = 1$). $N = N_{00} + N_{01} + N_{10} + N_{11}$ is the total number of sentences.

In this way, the score $\mathrm{I}(T = t; C = c)$ measures the relevance between word t and category c. Since we have multiple event factuality categories, for each word t, we take the average of these categorical scores, and finally we derive the overall relevance scores[3]:

$$\mathrm{Score}(t) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathrm{I}(T=t;C=c). \tag{6}$$

We measure the MI scores of words based on the training data, and select the top 30% words as the keywords, forming a factuality keyword set $\mathcal{K}$. Then, for each sentence, we select the words contained in $\mathcal{K}$ as the factuality-related keywords of the sentence.

[3] Following previous works [35, 4], we also focus on the category PS+, CT+ and CT-. We show examples of selected keywords in **Appendix**.

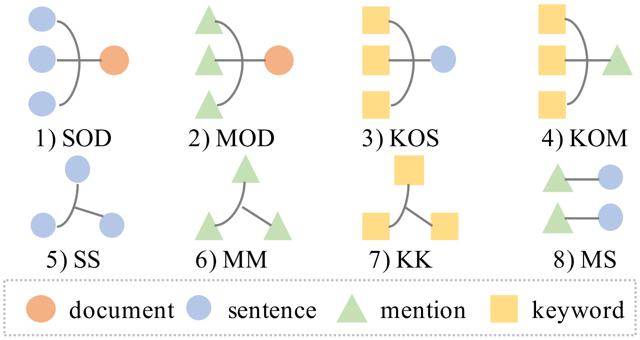

Figure 2. Relation set ($\mathcal{R}$) of hyperedges, where each one corresponds to an example of the hyperedge.

3.1.2 *Document Hypergraph Construction*

To promote the understanding of the factuality of events, we construct each document as a hypergraph. In summary, we mainly consider 4 types of nodes: ***document node***, ***sentence node***, ***mention node*** and ***keyword node***. Intuitively, the sentence/mention nodes retain local factuality features, and the keyword nodes provide valuable factuality cues associated with the sentences/mentions. The document node summarizes the global factuality features for prediction.

In order to establish complicated correlations in the document, we devise hyperedges with 8 relation types[4], termed the set $\mathcal{R}$, shown in Figure 2. Our main intuition is to encourage the global node (document) to consider different factuality features of local nodes (sentences/mentions/keywords), and the local nodes to consider factuality features with their related local nodes.

1. ***sentences-of-document (SOD)***: it connects the *document node* with all the contained *sentence nodes*.
2. ***mentions-of-document (MOD)***: it connects the *document node* with all the contained *mention nodes*.
3. ***keywords-of-sentence (KOS)***: it connects the *sentence node* with all the contained *keyword nodes*.
4. ***keywords-of-mention (KOM)***: it connects the *mention node* with all the *keyword nodes* associated in the same sentence.
5. ***sentences-sentences (SS)***: it connects all the *sentence nodes* contained in the document.
6. ***mentions-mentions (MM)***: it connects all the *mention nodes* referring to the same event in the document.
7. ***keywords-keywords (KK)***: it connects all the *keyword nodes* contained in the same sentence.
8. ***mention-sentence (MS)***: it connects the *mention node* with the corresponding *sentence node*.

Here, edges 1-2 establish the n-ary correlations between global and local nodes; edges 3-4 enhance the factuality features among local nodes; edge 8 preserves factuality features between mentions and their corresponding sentences; and edges 5-7 additionally emphasize factuality among the nodes of the same type.

[4] We perform extensive evaluations on the hyperedges and investigate the impact of the hypergraph structure in Table 3.

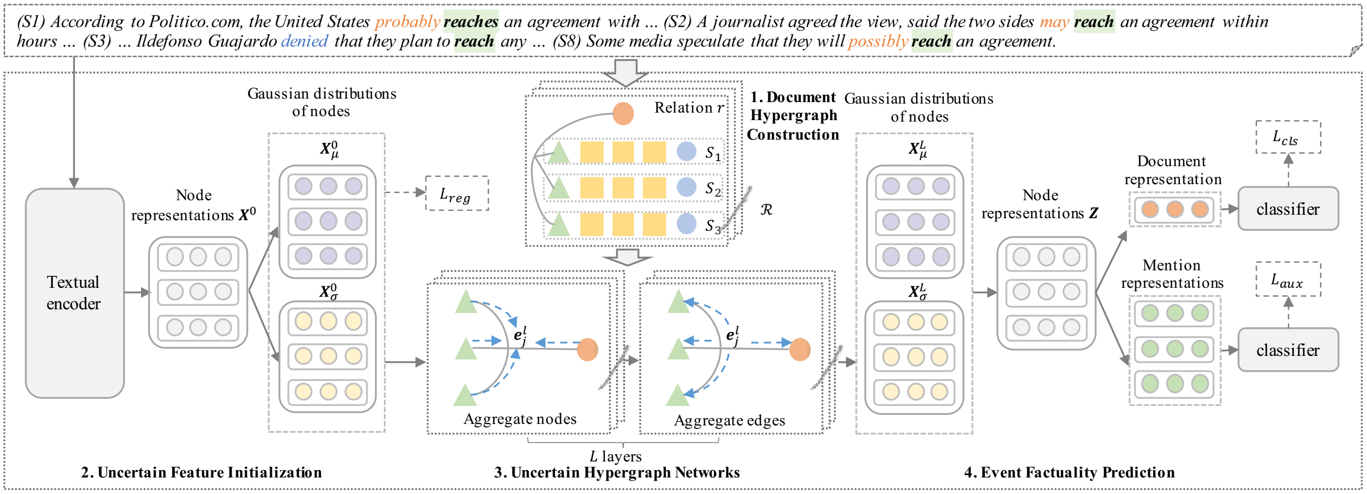

Figure 3. Overview of our proposed approach, UR-HAT. We first construct the document as a hypergraph. Then, we initialize uncertain node features with the textual encoder, perform uncertain hypergraph networks to refine node features, and finally predict factuality category with document-level understanding.

3.2 *Uncertain Feature Initialization*

To discern the different importance of factuality features, we represent node features using the uncertainty modeling paradigm. First, we derive certain textual features with BERT [8] for textual nodes. In general, BERT employs a transformer architecture [45], which generates textual features conditioned on contextual tokens and retains rich textual information. Formally, we input each sentence S_i of the given document into BERT to derive token features:

$$\boldsymbol{X}_i = \text{BERT}(S_i), i = 1, 2, ..., n_{\text{sent}}, \tag{7}$$

where n_{sent} is the number of sentences in the document. We take the feature of token `[CLS]` from $\boldsymbol{X}_i$, and collect all the sentences in the document to form sentence features $\boldsymbol{X}^{\text{sent}} \in \mathbb{R}^{d\times n_{\text{sent}}}$. Here d is the dimension of the BERT token features. We also derive the mention features $\boldsymbol{X}^{\text{men}} \in \mathbb{R}^{d\times n_{\text{men}}}$ and keyword features $\boldsymbol{X}^{\text{key}} \in \mathbb{R}^{d\times n_{\text{key}}}$ by applying mean pooling over the contained token features, where n_{men} and n_{key} is the total number of mentions and keywords in the current document, respectively. For the document feature $\boldsymbol{X}^{\text{doc}} \in \mathbb{R}^{d\times 1}$, we treat the document as a whole sentence [4], and input it into the same BERT to derive the `[CLS]` feature. Here, for the sake of symbol brevity, we concatenate all the above features as node features:

$$\boldsymbol{X}^0 = \text{concat}(\boldsymbol{X}^{\text{sent}}, \boldsymbol{X}^{\text{men}}, \boldsymbol{X}^{\text{key}}, \boldsymbol{X}^{\text{doc}}), \tag{8}$$

where $\boldsymbol{X}^0 \in \mathbb{R}^{d\times n}$, and $n = n_{\text{sent}} + n_{\text{men}} + n_{\text{key}} + 1$.

Having obtained the above textual features for the nodes, we estimate the uncertainty of the features. Specifically, we use a Gaussian distribution $\mathcal{N}(\boldsymbol{x}_\mu, \text{diag}(\boldsymbol{x}_\sigma))$ to represent each node's feature[5] rather than a deterministic feature. Here the mean $\boldsymbol{X}^0_\mu$ and variance $\boldsymbol{X}^0_\sigma$ of the node features can be derived by:

$$\boldsymbol{X}^0_\mu = \boldsymbol{W}_\mu \boldsymbol{X}^0, \quad \boldsymbol{X}^0_\sigma = \boldsymbol{W}_\delta \boldsymbol{X}^0, \tag{9}$$

where $\boldsymbol{W}_\mu$ and $\boldsymbol{W}_\delta$ are learnable parameters, used to generate means and variances. In this way, each node x_i is represented as a Gaussian distribution $\boldsymbol{x}^0_i = \mathcal{N}(\boldsymbol{x}^0_{\mu,i}, \text{diag}(\boldsymbol{x}^0_{\sigma,i}))$, which not only preserves the original feature about event factuality in the mean $\boldsymbol{x}^0_{\mu,i}$, but also estimates the uncertainty of the original feature in the variance $\boldsymbol{x}^0_{\sigma,i}$. These representations are further refined in the following networks.

3.3 *Uncertain Hypergraph Networks*

To refine factuality features with event correlations, we propose uncertain hypergraph networks upon document structures. For better illustration, this section will successively introduce the ***relation***, ***uncertainty*** and ***attention*** mechanisms to the general process in Eq. (3), and the final form of uncertain hypergraph networks is in Eq. (14).

Relation Mechanism. Considering that there are multiple relations (e.g., the relation set $|\mathcal{R}|$ is 8) in our document hypergraph, we first develop conventional HGCN [10] with the relation mechanism. Inspired by RGCN [41], we achieve the general process (Eq. (3)) in a relational multi-graph manner. Specifically, it aggregates node and hyperedge features with relation-specific parameters, which forms:

$$\begin{aligned} \boldsymbol{x}^{l+1}_i &= \delta\Big(\sum_{r\in\mathcal{R}}\sum_{e_j\in\mathcal{E}_i}\frac{1}{\bar{\boldsymbol{D}}_{r,jj}}\boldsymbol{e}^l_j\bar{\boldsymbol{W}}^l_r + \boldsymbol{x}^l_i\bar{\boldsymbol{W}}^l\Big), \\ \text{where } \boldsymbol{e}^l_j &= \delta\Big(\sum_{r\in\mathcal{R}}\sum_{v_k\in e_j}\frac{1}{\tilde{\boldsymbol{D}}_{r,kk}}\boldsymbol{x}^l_k\tilde{\boldsymbol{W}}^l_r\Big), \end{aligned} \tag{10}$$

where $\bar{\boldsymbol{D}}_r = \text{diag}(\sum_i \boldsymbol{H}_{r,ij}) \in \mathbb{R}^{m\times m}$ and $\tilde{\boldsymbol{D}}_r = \text{diag}(\sum_j \boldsymbol{H}_{r,ij}) \in \mathbb{R}^{n\times n}$ are the normalized incidence matrix $\boldsymbol{H}_r \in \{0,1\}^{n\times m}$ with the relation $r \in \mathcal{R}$. Note that $\boldsymbol{x}^{l+1}_i$ is the aggregated node feature and $\boldsymbol{e}^l_j$ is the aggregated hyperedge feature. $\bar{\boldsymbol{W}}^l_r$, $\tilde{\boldsymbol{W}}^l_r$ are learnable parameters with a specific relation r, and $\bar{\boldsymbol{W}}^l$ is the self-connection parameter.

Uncertainty Mechanism. As argued in previous work [54], modeling the node representation as a probability distribution is useful for estimating feature uncertainty. Therefore, we extend the deterministic relational hypergraph networks (Eq. (10)) in an uncertain manner. To illustrate, taking the node representation $\boldsymbol{x}_i = \mathcal{N}(\boldsymbol{x}_{\mu,i}, \text{diag}(\boldsymbol{x}_{\sigma,i}))$ as an example, we adopt the sum rule of prob-

[5] In this paper, we focus on the diagonal variance matrix as it is widely considered in studies [28, 54]. For the sake of brevity, we use $\boldsymbol{x}_\sigma$ to represent variances, instead of $\boldsymbol{x}^2_\sigma$.

ability distributions [28] for aggregation:

$$\sum_{i=1}^{n} w_i \boldsymbol{x}_i \sim \mathcal{N}(\sum_{i=1}^{n} w_i \boldsymbol{x}_{\mu,i}, \mathrm{diag}(\sum_{i=1}^{n} w_i^2 \boldsymbol{x}_{\sigma,i})), \tag{11}$$

where w_i stands for a constant scalar. Applying this rule to Eq. (10), we can achieve the node/hyperedge aggregation with uncertain features. Therefore, we perform the aggregation on the means and variances of the nodes and hyperedges, respectively, as follows:

$$\begin{aligned}
\boldsymbol{x}_{\mu,i}^{l+1} &= \delta(\sum_{r\in\mathcal{R}} \sum_{e_j\in\mathcal{E}_i} \frac{1}{\bar{\boldsymbol{D}}_{r,jj}} \boldsymbol{e}_{\mu,j}^{l} \bar{\boldsymbol{W}}_{\mu,r}^{l} + \boldsymbol{x}_{\mu,i}^{l} \bar{\boldsymbol{W}}_{\mu}^{l}),\\
\boldsymbol{x}_{\sigma,i}^{l+1} &= \delta(\sum_{r\in\mathcal{R}} \sum_{e_j\in\mathcal{E}_i} \frac{1}{\bar{\boldsymbol{D}}_{r,jj} \odot \bar{\boldsymbol{D}}_{r,jj}} \boldsymbol{e}_{\sigma,j}^{l} \bar{\boldsymbol{W}}_{\sigma,r}^{l} + \boldsymbol{x}_{\sigma,i}^{l} \bar{\boldsymbol{W}}_{\sigma}^{l}),\\
\boldsymbol{e}_{\mu,j}^{l} &= \delta(\sum_{r\in\mathcal{R}} \sum_{v_k\in e_j} \frac{1}{\tilde{\boldsymbol{D}}_{r,kk}} \boldsymbol{x}_{\mu,k}^{l} \tilde{\boldsymbol{W}}_{\mu,r}^{l}),\\
\boldsymbol{e}_{\sigma,j}^{l} &= \delta(\sum_{r\in\mathcal{R}} \sum_{v_k\in e_j} \frac{1}{\tilde{\boldsymbol{D}}_{r,kk} \odot \tilde{\boldsymbol{D}}_{r,kk}} \boldsymbol{x}_{\sigma,k}^{l} \tilde{\boldsymbol{W}}_{\sigma,r}^{l}),
\end{aligned} \tag{12}$$

where the derived node and hyperedge features can be written as $\boldsymbol{x}_i^{l+1} = \mathcal{N}(\boldsymbol{x}_{\mu,i}^{l+1}, \mathrm{diag}(\boldsymbol{x}_{\sigma,i}^{l+1}))$ and $\boldsymbol{e}_j^{l} = \mathcal{N}(\boldsymbol{e}_{\mu,j}^{l}, \mathrm{diag}(\boldsymbol{e}_{\sigma,j}^{l}))$, respectively. Here $\odot$ denotes element-wise production. For learning neural representations, we impose layer-specific parameters and non-linear activation functions on the means and variances, respectively, since they are mathematically intractable for distributions [54].

Attention Mechanism. Considering that nodes and hyperedges may have different importance, we further employ a variance-based attention mechanism to assign different weights to neighbors. Intuitively, a smaller variance implies a lower uncertainty of the feature, so we use a smooth exponential function to control the effect according to its variance as follows:

$$\boldsymbol{\alpha}_i = \exp(-\gamma \boldsymbol{x}_{\sigma,i}), \tag{13}$$

where γ is a scaling factor on the variance. Based upon this, we apply the attention mechanism to Eq. (12) and derive the entire **uncertain hypergraph networks** as follows:

$$\begin{aligned}
\boldsymbol{x}_{\mu,i}^{l+1} &= \delta(\sum_{r\in\mathcal{R}} \sum_{e_j\in\mathcal{E}_i} \frac{\boldsymbol{e}_{\mu,j}^{l} \odot \bar{\boldsymbol{\alpha}}_j^{l}}{\bar{\boldsymbol{D}}_{r,jj}} \bar{\boldsymbol{W}}_{\mu,r}^{l} + \boldsymbol{x}_{\mu,i}^{l} \bar{\boldsymbol{W}}_{\mu}^{l}),\\
\boldsymbol{x}_{\sigma,i}^{l+1} &= \delta(\sum_{r\in\mathcal{R}} \sum_{e_j\in\mathcal{E}_i} \frac{\boldsymbol{e}_{\sigma,j}^{l} \odot \bar{\boldsymbol{\alpha}}_j^{l} \odot \bar{\boldsymbol{\alpha}}_j^{l}}{\bar{\boldsymbol{D}}_{r,jj} \odot \bar{\boldsymbol{D}}_{r,jj}} \bar{\boldsymbol{W}}_{\sigma,r}^{l} + \boldsymbol{x}_{\sigma,i}^{l} \bar{\boldsymbol{W}}_{\sigma}^{l}),\\
\boldsymbol{e}_{\mu,j}^{l} &= \delta(\sum_{r\in\mathcal{R}} \sum_{v_k\in e_j} \frac{\boldsymbol{x}_{\mu,k}^{l} \odot \tilde{\boldsymbol{\alpha}}_k^{l}}{\tilde{\boldsymbol{D}}_{r,kk}} \tilde{\boldsymbol{W}}_{\mu,r}^{l}),\\
\boldsymbol{e}_{\sigma,j}^{l} &= \delta(\sum_{r\in\mathcal{R}} \sum_{v_k\in e_j} \frac{\boldsymbol{x}_{\sigma,k}^{l} \odot \tilde{\boldsymbol{\alpha}}_k^{l} \odot \tilde{\boldsymbol{\alpha}}_k^{l}}{\tilde{\boldsymbol{D}}_{r,kk} \odot \tilde{\boldsymbol{D}}_{r,kk}} \tilde{\boldsymbol{W}}_{\sigma,r}^{l}).
\end{aligned} \tag{14}$$

where $\bar{\boldsymbol{\alpha}}_j^l = \exp(-\gamma \boldsymbol{e}_{\sigma,j}^l)$ and $\tilde{\boldsymbol{\alpha}}_k^l = \exp(-\gamma \boldsymbol{x}_{\sigma,k}^l)$ measure the uncertainty of the hyperedge and node features, respectively. In this way, the generated node features take into account the feature uncertainty of the related hyperedges, while the generated hyperedge features take into account the feature uncertainty of the related nodes, thus helping to aggregate uncertain local features in both node and hyperedge aspects. In the following section, we denote the final output node features as $\boldsymbol{x}_{\mu,i}$ and $\boldsymbol{x}_{\sigma,i}$ by omitting the superscript l.

3.4 Event Factuality Prediction

Based on the learned uncertain node features $\mathcal{N}(\boldsymbol{x}_{\mu,i}, \mathrm{diag}(\boldsymbol{x}_{\sigma,i}))$, we generate deterministic features and then make predictions.

3.4.1 Prediction Layer

Since the distributions are intractable for classification [54], we perform a sampling process for all the node features, which is $\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{x}_{\mu,i}, \mathrm{diag}(\boldsymbol{x}_{\sigma,i}))$. However, such a direct sampling process is intractable for back-propagation, so we use the well-known reparameterization trick [16] as a solution. Mathematically, we first sample $\boldsymbol{\epsilon}$ from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and then transform it into the target distribution:

$$\boldsymbol{z}_i = \boldsymbol{x}_{\mu,i} + \boldsymbol{\epsilon} \odot \sqrt{\boldsymbol{x}_{\sigma,i}}, \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}). \tag{15}$$

Afterward, we derive the deterministic representation of the document node $\boldsymbol{z}^{\text{doc}}$ and the mention nodes $\boldsymbol{z}_j^{\text{men}}$, and then use them to predict the event factuality categories of the document and mentions:

$$\begin{aligned}
\boldsymbol{p}^{\text{doc}} &= \mathrm{softmax}(\boldsymbol{W}^{\text{doc}} \boldsymbol{z}^{\text{doc}} + \boldsymbol{b}^{\text{doc}}),\\
\boldsymbol{p}_j^{\text{men}} &= \mathrm{softmax}(\boldsymbol{W}^{\text{men}} \boldsymbol{z}_j^{\text{men}} + \boldsymbol{b}^{\text{men}}),
\end{aligned} \tag{16}$$

Note that our goal is to predict event factuality $\boldsymbol{p}^{\text{doc}}$ at the document level as output, and here we additionally predict the event factuality $\boldsymbol{p}_j^{\text{men}}$ of mentions as an auxiliary supervision in the training phase.

3.4.2 Training Objective

For training, we adopt the cross-entropy loss on the event factuality predictions of both the document and the mentions:

$$\begin{aligned}
\mathcal{L}_{cls} &= -\boldsymbol{y}^{\text{doc}} \cdot \log(\boldsymbol{p}^{\text{doc}}),\\
\mathcal{L}_{aux} &= -\sum_{j=1}^{n_{\text{men}}} \boldsymbol{y}_j^{\text{men}} \cdot \log(\boldsymbol{p}_j^{\text{men}}),
\end{aligned} \tag{17}$$

The $\mathcal{L}_{cls}$ is used to achieve document-level event factuality prediction, and the $\mathcal{L}_{aux}$ is used to refine mention factuality features[6]. Furthermore, we design a regulation loss to encourage the node distributions to approximate the Gaussian distribution, achieved by:

$$\mathcal{L}_{reg} = \mathrm{KL}(\mathcal{N}(\boldsymbol{X}_{\mu}^{0}, \boldsymbol{X}_{\sigma}^{0}) || \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})), \tag{18}$$

where $\mathcal{N}(\boldsymbol{X}_{\mu}^{0}, \boldsymbol{X}_{\sigma}^{0})$ is the Gaussian distributions for all node features, and $\mathrm{KL}(\cdot||\cdot)$ is the KL-divergence [19] between distributions. Finally, the overall objective of a document is:

$$\mathcal{L} = \mathcal{L}_{cls} + \beta_1 \mathcal{L}_{aux} + \beta_2 \mathcal{L}_{reg}, \tag{19}$$

where β_1 and β_2 are harmonic factors. We optimize all the training documents in the mini-batch strategy with AdamW [23].

4 Experiment

In this section, we compare our proposed model with previous models, and conduct experiments to evaluate our model.

4.1 Dataset and Evaluation Metric

We conduct experiments on two widely used datasets, English and Chinese event factuality datasets [35], which contain 1,730 and 4,650 documents, respectively. Following previous works [35, 4], since the PS- and Uu categories cover only 1.39% and 1.20% of the documents in the English and Chinese datasets, we also mainly focus on the performance of the CT+, CT- and PS+ categories. See **Appendix** for more details. For a fair comparison, we adopt 10-fold cross-validation with exactly the same data split as previous works. For evaluation, we adopt F1 scores with both the micro- and macro-averaged metrics for the overall performance.

[6] Note that the event factuality of each mention is provided by the datasets, and we only use it for refining features in the training phase.

Table 1. 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 ULGN. The Wilcoxons test shows significant difference (p<0.05) on both F1 metrics.

Datasets	Methods	CT+	CT-	PS+	Micro-F1	Macro-F1
English	MaxEntVote [35]	75.14	58.17	35.89	68.42	56.40
	BiLSTM-Att [35]	79.18	65.25	53.65	73.23	66.03
	SentVote [35]	83.98	70.22	57.85	78.06	70.68
	Att-Adv [35]	89.84	76.87	62.14	83.56	76.28
	BERT Model [8]	89.38	71.82	69.09	83.53	76.76
	GCNN [51]	91.19	80.28	70.76	86.37	80.74
	ULGN [4]	92.49	84.87	76.68	88.69	84.68
	UR-HAT (Ours)	**93.98** (↑ 1.49)	**89.25** (↑ 4.38)	**78.87** (↑ 2.19)	**90.93** (↑ 2.24)	**87.37** (↑ 2.69)
Chinese	MaxEntVote [35]	72.22	62.44	58.29	67.72	64.32
	BiLSTM-Att [35]	81.89	68.82	49.78	71.12	67.28
	SentVote [35]	80.68	72.66	58.39	74.70	70.58
	Att-Adv [35]	87.52	83.35	74.06	84.03	81.64
	BERT Model [8]	84.79	88.71	79.33	85.83	84.28
	GCNN [51]	89.60	85.38	76.81	86.03	83.93
	ULGN [4]	93.53	94.99	**90.76**	93.77	93.09
	UR-HAT (Ours)	**94.14** (↑ 0.61)	**95.44** (↑ 0.45)	90.51 (↓ 0.25)	**94.26** (↑ 0.49)	**93.36** (↑ 0.27)

4.2 Implementation Details

For the implementation, we adopt BERT-base model [8] as the textual encoder. The initial learning rate is tuned in {2e-5, 3e-5, 5e-5} with linear decay. The batch size is tuned in $\{1, 2, 3, 4\}$ for both datasets. The number of graph layers is tuned in $\{1, 2, ..., 6\}$. The harmonic factor β_1 and β_2 is tuned in $[0, 1]$ and [1e-4,1e-3], respectively. The scaling factor γ is tuned in [1e-4,1e-2]. The optimal hyperparameters are tuned on the validation set by grid search according to the averaged micro and macro F1 scores, which can be found in **Appendix**. For baseline methods, we copy the official results from their original literature to avoid re-implementation bias.

Table 2. Results (%) on variants of the uncertain hypergraph networks.

Datasets	Methods	Micro-F1	Macro-F1	ΔAvg
English	Entire Model	**90.93**	**87.37**	-
	w/o auxiliary loss	89.94	85.75	↓ 1.31
	w/o attention	89.98	86.04	↓ 1.14
	w/o uncertainty	89.48	86.27	↓ 1.28
	w/o relation	89.42	84.83	↓ 2.03
	w/o hypergraph	86.05	79.44	↓ 6.41
	repl. HGCN	89.21	85.52	↓ 1.79
	repl. ULGN	88.52	84.02	↓ 2.88
	repl. GCN	88.22	84.21	↓ 2.94
	BERT Model	83.53	76.76	↓ 9.01

4.3 Baselines

For a fair comparison[7], we choose the following methods as baselines: (1) **MaxEntVote** [35], which employs the maximum entropy model to identify sentence-level event factuality, and then votes the factuality with the most sentences as the final prediction. (2) **BiLSTM-Att** [35], which employs BiLSTM for feature extraction, and uses intra-sentence attention to capture important information in the sentence. (3) **SentVote** [35], which leverages the intra-sentence attention to identify sentence-level event factuality, and adopts a voting strategy to make the document-level prediction. (4) **Att-Adv** [35], which leverages the intra- and inter-sentence attention to learn the document representation, and utilizes adversarial training to improve the robustness. (5) **BERT Model**, which utilizes the vanilla BERT-base model [8] to encode documents, and uses the `[CLS]` representation for prediction, treating the document as a long sentence. (6) **GCNN** [51], which adopts gated convolutional neural networks with self-attention layers, and utilizes additional syntactic features to identify the factuality. (6) **ULGN** [4], which deploys the document as a conventional homogeneous graph, and adopts uncertain local-to-global graph networks to aggregate local sentence-level event factuality features for the document-level event factuality prediction.

[7] We investigate recent studies [36, 37, 53], but they all require external resources (such as cross-domain event data or additional event description texts), which have quite different task settings for a fair comparison.

4.4 Main Results

The results on both datasets are shown in Table 1, which shows that: (1) ***Our model outperforms all the baselines, indicating the effectiveness of our model.*** For example, our model outperforms the best traditional baseline and the BERT-based baseline on the English data by 7.37 and 2.24 micro-F1 improvements respectively, reflecting the effectiveness of the model. (2) ***The hypergraph structure can capture more meaningful information for the DocEFI task.*** Compared to the previous state-of-the-art method ULGN, our model achieves better results. Considering that ULGN is developed on the conventional graph, our hypergraph structure can naturally capture n-ary correlations, which provides a better capability for event factuality inference. (3) ***Our model is general on the two benchmarks in different languages.*** Most of the results are higher on Chinese data than on English data, and we find that the scale of training data is larger on the Chinese data, where the model can learn richer features. However, our model consistently outperforms previous baselines, especially on the relatively smaller English data, which also reflects the effectiveness of our model. In addition, we find that the PS+ category derives relatively lower results on the Chinese data, and we find an imbalance in the data across categories: CT-/1342, CT+/2403 and PS+/848 of category/documents. We believe that the relatively small data scale makes it difficult for the model to fully learn the ambiguous PS+ category. Considering the significant F1 improvements on both datasets, our model is worthy for DocEFI.

Table 3. Results (%) on variants of the constructed document hypergraph.

Datasets	Methods	Micro-F1	Macro-F1	ΔAvg
English	Entire Graph (ours)	**90.93**	**87.37**	-
	w/o node: mention	87.95	83.67	↓ 3.34
	w/o node: sentence	90.23	86.46	↓ 0.81
	w/o node: keyword	89.80	85.61	↓ 1.45
	w/o edge: MOD	87.99	83.82	↓ 3.25
	w/o edge: SOD	89.88	85.91	↓ 1.26
	w/o edge: KOM	89.48	86.05	↓ 1.39
	w/o edge: KOS	89.73	86.33	↓ 1.12
	w/o edge: MS	90.39	86.65	↓ 0.63
	w/o edge: KK,MM,SS	90.19	86.47	↓ 0.82
	w/o hypergraph	86.05	79.44	↓ 6.41

4.5 Analysis on Hypergraph Networks

To investigate the impact of each mechanism in the model, we run variants on the proposed hypergraph neural networks. The results are shown in Table 2, which shows that: (1) ***The "attention", "uncertainty", and "relation" mechanisms have a positive impact on the results.*** The attention mechanism helps to detect valuable node features with higher confidence. The uncertainty mechanism measures the uncertainty of local features across sentences, helping to resolve the conflicting event factuality with the degree of feature uncertainty. The relation mechanism helps to capture different semantic correlations in the graph, allowing feature aggregation in different ways. (2) ***The proposed hypergraph model can be more beneficial for the DocEFI task than previous typical graph models.*** *"repl. HGCN"* replaces the proposed model with vanilla HGCN. *"repl. ULGN"* and *"repl. GCN"* treat hyperedges as fully connected subgraphs, and on the graph adopt ULGN [4] and GCN respectively. It reflects the hypergraph is more inferential for DocEFI, and the uncertainty manner captures valuable information for event factuality. We provide more details and results on both datasets in **Appendix**.

4.6 Analysis on Hypergraph Structures

To investigate the effect of each sub-structure in the hypergraph, we run variants on the hypergraph structure. The results are shown in Table 3, which shows that (1) ***All nodes retain useful information for the results.*** In particular, the mention nodes retain crucial local event information. The keyword nodes highlight fine-grained factuality cues for the event in each sentence. The sentence nodes preserve contextual information, which not only preserves factuality cues, but also preserves syntactic and semantic meanings in texts. (2) ***All edges provide useful semantic connections, especially the edges associated with the document node.*** For example, the *"MOD"* and *"SOD"* edges establish the connections between document-level and sentence-level nodes, which helps to aggregate valuable information from local sentences to the global document. Furthermore, the *"KOM"* and *"KOS"* edges establish the connections between sentence-level and word-level nodes, which aggregates information from local keywords to the sentences involved. The *"MS"* edges further represent the specific connections between the mentions and the sentences, providing the corresponding relationship. The additional *"KK,MM,SS"* edges enclose the nodes with similar roles, also providing useful effects. (3) ***The document hypergraph explicitly strengthens the correlations among the related textual components.*** *"w/o hypergraph"* removes the hypergraph structures but remain uncertainty mechanism on node representations. The hypergraph serves as prior knowledge of the textual structures, so we can observe the results decrease significantly without document hypergraph. We provide more details and results on both datasets in **Appendix**.

Table 4. Results (%) on the document groups of n_c categories of sentence-level factuality in the English dataset.

n_c	Methods	Micro-F1	Macro-F1
$n_c = 1$	Att-Adv	91.36	81.67
	ULGN	92.48	85.21
	UR-HAT	**94.72** (↑ 2.24)	**88.07** (↑ 2.86)
$n_c > 1$	Att-Adv	60.91	60.04
	ULGN	75.51	74.76
	UR-HAT	**76.73** (↑ 1.22)	**76.22** (↑ 1.46)

4.7 Analysis on Factuality Category Number

Considering there exist different sentence-level event factuality in the document, we investigate our model on different groups of documents. The results are shown in Table 4, which shows that (1) ***In both data groups, our model achieves higher performance than previous models.*** We attribute the reason to the fact that the hypergraph can be inferential for document understanding, thus improving the document results, especially for the $n_c > 1$ group. In addition, our model exploits valuable factuality-related keywords, benefiting both groups. (2) ***Also, all methods achieve lower results for the $n_c > 1$ group, indicating the need to integrate local factuality conflicts.*** Our model still achieves better results, reflecting the advantage of modeling the factuality conflicts of the sentence-level event factuality.

5 Related Work

Event factuality identification (EFI) is an important task in information extraction, which can be helpful for event extraction [6, 24, 43, 44] and knowledge acquisition [42, 21, 22]. Pioneering studies explore the ***sentence-level EFI task***, which have explored rule-based methods [26, 39], machine learning methods [30, 40, 20, 33] and neural networks [38, 34, 14, 29]. However, an event can have conflicting factuality in different sentences, usually leading to confusing factuality results in applications. Recent studies [35, 4, 36, 37, 49] have attempted the ***document-level EFI task***, which aims to identify event factuality at the document level. They explore attention mechanisms [35, 36] or graph convolutional networks [4, 52] to aggregate sentence factuality. However, they either neglect the global graph structures in the document or simply represent the document as a homogeneous graph, neglecting the crucial n-ary correlations. There are also studies [36, 37, 53] that use external resources (like cross-domain data) to enrich information, which is quite different from our task. In this paper, we extend previous studies with hypergraphs to exploit n-ary correlations in document structures, and estimate feature uncertainty to exploit different local event factuality.

Hypergraph neural networks are proven effective in several graph-based applications [9]. Traditional graph neural networks [17, 46, 41] mostly focus on the typical graphs with pairwise edges. In order to capture n-ary correlations with multiple nodes, recent studies [10, 1, 13] investigate hypergraph neural networks. In general, there are several spectral-based studies [10, 2] and spatial-based studies [1, 13]. Among them, HGCN [10] develops GCN [17] into hypergraphs with graph convolutions. UniGNN [13] further generalizes it with several classical GNN layer forms. There are also studies [9] applying hypergraphs in NLP tasks. In this paper, we use hypergraphs to exploit the n-ary correlations in document structures to promote the understanding of event factuality at the document level.

Uncertainty modeling is a crucial topic in deep learning studies [54]. Traditional deep learning methods usually represent features as deterministic vectors [32], which can hardly express the uncertainty of the features [11, 15]. To this end, variational autoencoder (VAE) [16] represents features as probability distributions, providing an effective way to capture the uncertainty of data. Afterwards, Vilnis and McCallum [47] propose Gaussian embeddings for word representations considering word ambiguity. He et al. [12] introduce uncertainty modeling for entity and relation representations for knowledge graph embeddings. There are also studies [50] investigate uncertainty in few-shot scenarios for knowledge graph completion. Xiao and Wang [48] review previous works and explore uncertainty in various NLP tasks. In this paper, we explore uncertainty of event factuality features and propose an uncertain neural form for hypergraph representation learning.

6 Conclusion

This paper proposes a novel approach using uncertain relational hypergraph attention networks, namely UR-HAT, for the DocEFI task. Specifically, we reframe the document graph as a hypergraph and establish the n-ary correlations in the document. To better capture the importance of event factuality features, we further encode textual node features with uncertain Gaussian distributions, and propose a novel uncertain hypergraph network on top of the uncertain node features to summarize document-level event factuality. Experiments indicate the effectiveness on two widely-used datasets. Our future work will enhance document structures, evaluate uncertain hypergraph networks in general tasks, and adapt large language models for this downstream task.

Acknowledgements

Corresponding authors: Lihong Wang and Shu Guo. We would like to thank all reviewers for their insightful comments. This work is supported by the National Key Research and Development Program of China (Grant No.2021YFB3100600), the National Natural Science Foundation of China (No.62106059), the National Social Science Foundation of China (Grant No.19BSH022), and the Youth Innovation Promotion Association of CAS (Grant No.2021153).

References

[1] Devanshu Arya, Deepak K. Gupta, Stevan Rudinac, and Marcel Worring. Hypersage: Generalizing inductive representation learning on hypergraphs. *ArXiv*, abs/2010.04558, 2020.

[2] Song Bai, Feihu Zhang, and Philip H. S. Torr. Hypergraph convolution and hypergraph attention. *Pattern Recognit.*, 2021.

[3] Alain Bretto. Hypergraph theory: An introduction. 2013.

[4] Pengfei Cao, Yubo Chen, Yuqing Yang, Kang Liu, and Jun Zhao. Uncertain local-to-global networks for document-level event factuality identification. In *Proceedings of EMNLP*, 2021.

[5] Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, Yuguang Chen, and Weihua Peng. Knowledge-enriched event causality identification via latent structure induction networks. In *Proceedings of ACL*, 2021.

[6] Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of ACL*, 2015.

[7] Shiyao Cui, Jiawei Sheng, Xin Cong, Quangang Li, Tingwen Liu, and Jinqiao Shi. Event causality extraction with event argument correlations. In *Proceedings of COLING*, pages 2300–2312, 2022.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, 2019.

[9] Kaize Ding, Jianling Wang, Jundong Li, Dingcheng Li, and Huan Liu. Be more with less: Hypergraph attention networks for inductive text classification. In *Proceedings of EMNLP*, 2020.

[10] Yifan Feng, Haoxuan You, Zizhao Zhang, R. Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of AAAI*, 2019.

[11] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of ICML*, 2016.

[12] Shizhu He, Kang Liu, Guoliang Ji, and Jun Zhao. Learning to represent knowledge graphs with gaussian embedding. In *Proceedings of CIKM*, 2015.

[13] Jing Huang and Jie Yang. Unignn: a unified framework for graph and hypergraph neural networks. In *Proceedings of IJCAI*, 2021.

[14] Rongtao Huang, Bowei Zou, Hongling Wang, Peifeng Li, and Guodong Zhou. Event factuality detection in discourse. In *Proceedings of NLPCC*, 2019.

[15] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of NeurIPS*, 2017.

[16] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of ICLR*, 2014.

[17] Thomas Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of ICLR*, 2017.

[18] Manfred Klenner and Simon Clematide. How factuality determines sentiment inferences. In *Proceedings of *SEM*, 2016.

[19] Solomon Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 1951.

[20] Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. Event detection and factuality assessment with non-expert supervision. In *Proceedings of EMNLP*, 2015.

[21] Qian Li, Shu Guo, Cheng Ji, Xutan Peng, Shiyao Cui, Jianxin Li, and Lihong Wang. Dual-gated fusion with prefix-tuning for multi-modal relation extraction. In *Findings of ACL*, pages 8982–8994, 2023.

[22] Qian Li, Shu Guo, Yangyifei Luo, Cheng Ji, Lihong Wang, Jiawei Sheng, and Jianxin Li. Attribute-consistent knowledge graph representation learning for multi-modal entity alignment. In *Proceedings of WWW*, pages 2499–2508, 2023.

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of ICLR*, 2019.

[24] Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang, and Jing Shao. Prompt for extraction? PAIE: Prompting argument interaction for event argument extraction. In *Proceedings of ACL*, 2022.

[25] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. 2005.

[26] Rowan Nairn, Cleo Condoravdi, and Lauri Karttunen. Computing relative polarity for textual inference. In *Proceedings of ICoS-5*, 2006.

[27] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

[28] Valentin V. Petrov. Sums of independent random variables. 1975.

[29] Amir Pouran Ben Veyseh, Thien Huu Nguyen, and Dejing Dou. Graph based neural networks for event factuality prediction using syntactic and semantic structures. In *Proceedings of ACL*, 2019.

[30] Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. Automatic committed belief tagging. In *Proceedings of COLING*, 2010.

[31] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of EMNLP*, 2011.

[32] Chen Qian, Fuli Feng, Lijie Wen, and Tat-Seng Chua. Conceptualized and contextualized gaussian embedding. In *Proceedings of AAAI*, 2021.

[33] Zhong Qian, Peifeng Li, and Qiaoming Zhu. A two-step approach for event factuality identification. In *Proceedings of IALP*, 2015.

[34] Zhong Qian, Peifeng Li, Yue Zhang, Guodong Zhou, and Qiaoming Zhu. Event factuality identification via generative adversarial networks with auxiliary classification. In *Proceedings of IJCAI*, 2018.

[35] Zhong Qian, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. Document-level event factuality identification via adversarial neural network. In *Proceedings of NAACL*, 2019.

[36] Zhong Qian, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. Document-level event factuality identification via reinforced multi-granularity hierarchical attention networks. In *Proceedings of IJCAI*, 2022.

[37] Zhong Qian, Heng Zhang, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. Document-level event factuality identification via machine reading comprehension frameworks with transfer learning. In *Proceedings of COLING*, 2022.

[38] Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. Neural models of factuality. In *Proceedings of NAACL*, 2018.

[39] Roser Saurí. A factuality profiler for eventualities in text. *Brandeis University*, 2008.

[40] Roser Saurí and James Pustejovsky. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics*, 2012.

[41] M. Schlichtkrull, Thomas Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. *ArXiv*, 2018.

[42] Jiawei Sheng, Shu Guo, Zhenyu Chen, Juwei Yue, Lihong Wang, Tingwen Liu, and Hongbo Xu. Adaptive attentional network for few-shot knowledge graph completion. In *Proceedings of EMNLP*, pages 1681–1691, 2020.

[43] Jiawei Sheng, Shu Guo, Bowen Yu, Qian Li, Yiming Hei, Lihong Wang, Tingwen Liu, and Hongbo Xu. Casee: A joint learning framework with cascade decoding for overlapping event extraction. In *Findings of ACL*, pages 164–174, 2021.

[44] Jiawei Sheng, Rui Sun, Shu Guo, Shiyao Cui, Jiangxia Cao, Lihong Wang, Tingwen Liu, and Hongbo Xu. Cored: Incorporating type-level and instance-level correlations for fine-grained event detection. In *Proceedings of SIGIR*, pages 1122–1132, 2022.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of NeurIPS*, 2017.

[46] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *Proceedings of ICLR*, 2018.

[47] Luke Vilnis and Andrew McCallum. Word representations via gaussian embedding. *CoRR*, abs/1412.6623, 2015.

[48] Yijun Xiao and William Yang Wang. Quantifying uncertainties in natural language processing tasks. In *Proceedings of AAAI*, 2018.

[49] Heng Zhang, Zhong Qian, Peifeng Li, and Xiaoxu Zhu. Evidence-based document-level event factuality identification. In *Proceedings of PRICAI*, 2022.

[50] Jiatao Zhang, Tianxing Wu, and Guilin Qi. Gaussian metric learning for few-shot uncertain knowledge graph completion. In *Proceedings of DASFAA*, 2021.

[51] Yun Zhang, Peifeng Li, and Qiaoming Zhu. Document-level event factuality identification method with gated convolution networks. *Computer Science*, 2020.

[52] Zihao Zhang, Chengwei Liu, Zhong Qian, Xiaoxu Zhu, and Peifeng Li. Hs^{2}n: Heterogeneous semantics-syntax fusion network for document-level event factuality identification. In *Proceedings of PRICAI*, 2022.

[53] Zihao Zhang, Zhong Qian, Xiaoxu Zhu, and Peifeng Li. Code: Contrastive learning method for document-level event factuality identification. In *Proceedings of DASFAA*, 2023.

[54] Dingyuan Zhu, Ziwei Zhang, Peng Cui, and Wenwu Zhu. Robust graph convolutional networks against adversarial attacks. In *Proceedings of SIGKDD*, 2019.