Cross-document Misinformation Detection based on Event Graph Reasoning

Anonymous ACL submission

Abstract

For emerging events, human readers are often exposed to both real news and fake news. Multiple news articles may contain complementary or contradictory information that readers can leverage to help detect fake news. Inspired by this process, we propose a novel task of cross-document misinformation detection. Given a cluster of topically related news documents, we aim to detect misinformation at both document level and a more fine-grained level, event level. Due to the lack of data, we generate fake news by manipulating real news, and construct 3 new datasets with 422, 276, and 1,413 clusters of topically related documents, respectively. We further propose a graph-based detector that constructs a cross-document knowledge graph using cross-document event coreference resolution and employs a heterogeneous graph neural network to conduct detection at two levels. We then feed the event-level detection results into the document-level detector. Experimental results show that our proposed method significantly outperforms existing methods by up to 7 F1 points on this new task.¹

1 Introduction

The dissemination of fake news has become an important social issue. For emergent complex events, human readers are usually exposed to multiple news documents, where some are real and others are fake. News documents from different sources naturally form a cluster of topically related documents. We notice that articles about the same topic may contain conflicting or complementary information, which can benefit the task of misinformation detection. An example is shown in Figure 1. As shown in the knowledge graph, the death of Rosanne Boyland in 2021 US Capitol attack is a shared event across all four documents. Each document is internally consistent, which makes it difficult to identify misinformation when judging each news separately. However, the three real news documents complement each other’s statements regarding the death of Boyland, while the fake news document contradicts the other stories. Such cross-document connections can be leveraged to help detect misinformation.

Most existing work in fake news detection is limited to judging each document in isolation. In contrast, we propose a novel task of cross-document misinformation detection that aims to detect fake information from a cluster of topically related news documents. We conduct the task at both document level and event level. Each event describes a specific type of real-world event mentioned in the text (e.g., the death of Boyland in Figure 1), and usually involves certain participants to represent different aspects of the event (e.g., the death cause and the victim of the death event). Document-level detection aims to detect fake news documents. Event-level detection is a more fine-grained task that aims to detect fake events, thereby pinpointing specific fake information in news documents.

Existing work on fine-grained misinformation detection detects fake knowledge triplets (Fung et al., 2021). However, we focus on identifying false events instead of relations or entities, because events are more important to storytelling, and easier to compare across multiple documents through cross-document coreference resolution.

To the best of our knowledge, there are no fake news detection datasets with clusters of topically related documents. Therefore, we construct 3 new benchmark datasets based on existing real news corpus with such clusters. Following Fung et al. (2021), we train a generator that generates a document from a knowledge graph (KG), and feed manipulated KGs into the generator to generate fake news documents. By tracking the manipulation operations, we also obtain supervision for

¹We attach the codes in the submission and upload the datasets to google drive at LINK.
Figure 1: An example of cross-document misinformation detection, including the texts and knowledge graphs for four news documents. The three real news documents complement each other, while the fake news contradicts the other news. News 1 falsely speculates that Boyland was crushed to death, but it admits that the cause of death was not yet verified. News 2 and 3 complete the story by reporting that Boyland died of drug overdose. The fake news claims that Boyland was killed by police, which contradicts the other news. Additionally, the fake news states that the police attacked Boyland, which is inconsistent with News 3’s claim that the police was trying to help her.

event-level detection.

We further propose a detection system as shown in Figure 2. Given a cluster of documents, we first use an IE system (Lin et al., 2020) to construct a within-document KG for each document. Then, we connect the within-document KGs to form a cross-document KG using cross-document event coreference resolution (Lai et al., 2021; Wen et al., 2021). Eventually, we use a heterogeneous graph neural network (GNN) to encode the cross-document KG and conduct detection at two levels.

Our contributions are summarized as follows:
1. We propose the novel task of cross-document misinformation detection, and conduct the task at two levels, document level and the more fine-grained event level.
2. We construct 3 new datasets for our proposed task based on existing document clusters categorized by topics.
3. We propose a detector that leverages cross-document information and improve document-level detection by utilizing features produced by the event-level detector. Experiments on 3 datasets demonstrate that our method significantly outperforms existing methods.

2 Related Work

Fake News Detection: Early work for fake news detection uses hand-crafted features to conduct document classification (Rubin et al., 2016; Wang, 2017; Rashkin et al., 2017; Pérez-Rosas et al., 2018; Sarkar et al., 2018; Atanasova et al., 2019). Recent work uses neural network such as RNN (Karimi et al., 2018; Nasir et al., 2021) and Transformer (Zellers et al., 2019) to encode the news document. To model the internal structure of a news document, Karimi and Tang (2019) models the inter-sentence dependency tree, Vaibhav et al. (2019) and Hu et al. (2021) model the interactions between sentences, and Pan et al. (2018) and Fung et al. (2021) model the knowledge graph extracted by IE systems. Similar to our work, Hu et al. (2021) compares the news with external knowledge base (KB) to check for inconsistencies. However, the correlation between news and KB is not as close as the correlation between related news documents due to the incompleteness of these KBs. Other work utilizes additional information such as user engagements and behaviors on social media (Shu et al., 2019; Nguyen et al., 2020) and multi-modal features (Khattar et al., 2019; Tan et al., 2020; Fung et al., 2021). However, to the best of our knowledge, no published work has considered using cross-document inference for misinformation detection.

In addition to document-level detection, the task of fine-grained detection is also important but rarely explored. The most relevant work detects fake knowledge triplets extracted from each individual news article (Fung et al., 2021).

Another related task is fact verification which aims to verify a statement based on retrieved evidence. Fact verification has been explored in multiple domains such as general domain (Thorne et al., 2018), climate change (Diggelmann et al., 2020) and COVID-19 (Wadden et al., 2020; Saakyan et al., 2018; Nasir et al., 2021) and Transformer (Zellers et al., 2019) to encode the news document. To model the internal structure of a news document, Karimi and Tang (2019) models the inter-sentence dependency tree, Vaibhav et al. (2019) and Hu et al. (2021) model the interactions between sentences, and Pan et al. (2018) and Fung et al. (2021) model the knowledge graph extracted by IE systems. Similar to our work, Hu et al. (2021) compares the news with external knowledge base (KB) to check for inconsistencies. However, the correlation between news and KB is not as close as the correlation between related news documents due to the incompleteness of these KBs. Other work utilizes additional information such as user engagements and behaviors on social media (Shu et al., 2019; Nguyen et al., 2020) and multi-modal features (Khattar et al., 2019; Tan et al., 2020; Fung et al., 2021). However, to the best of our knowledge, no published work has considered using cross-document inference for misinformation detection.

In addition to document-level detection, the task of fine-grained detection is also important but rarely explored. The most relevant work detects fake knowledge triplets extracted from each individual news article (Fung et al., 2021).

Another related task is fact verification which aims to verify a statement based on retrieved evidence. Fact verification has been explored in multiple domains such as general domain (Thorne et al., 2018), climate change (Diggelmann et al., 2020) and COVID-19 (Wadden et al., 2020; Saakyan...
et al., 2021). However, fact verification focuses on short single-sentence statements, and cannot model the complicated internal structure of a news document.

**Fake News Datasets:** The main difficulty in constructing a fake news dataset is to obtain annotations. Rashkin et al. (2017) and Rubin et al. (2016) obtain labels from the source information, and consider news from reliable sources as real news, and unreliable sources as fake news. A potential issue is that the detector may only learn to distinguish the style of different news sources, rather than the authenticity of the content. Shu et al. (2020) collects annotations from fact-checking websites, and Pérez-Rosas et al. (2018) collects annotations via crowd-sourcing. These approaches produce datasets of higher quality, but require extensive manual efforts. With the development of powerful generative models capable of mimicking human-written news (Zellers et al., 2019), recent work has constructed datasets by using generative models to generate fake news (Tan et al., 2020; Fung et al., 2021). Fung et al. (2021) further generates fake news from manipulated KG, which we follow to construct our dataset.

### 3 Task Formulation

Given a cluster of documents about the same story, the task of cross-document misinformation detection aims to detect the fake information included in the cluster.

Formally, let \( S = \{d_1, \ldots, d_N\} \) be the document cluster, and \( N = |S| \) be the size of the cluster. Some documents in \( S \) are real, while others are fake. From each document \( d \in S \), we extract events \( E(d) = \{e_1, \ldots, e_m\} \), where \( m = |E(d)| \) is the number of events in document \( d \). In an extracted event set \( E(d) \), some events are real and others are fake.

We conduct the task of misinformation detection at two levels, document level and event level. **Document-level detection** aims to predict whether each document \( d \in S \) is real or fake. **Event-level detection** is a more fine-grained task which aims to predict whether each event \( e \in E(d), d \in S \) is real or fake. In the example in Figure 1, the die event in the fake news is fake, since it falsely describes Boyland being killed by the police but she actually died of drug overdose.

### 4 Approach

An overview of our approach is shown in Figure 2. Given a cluster of documents, we first construct a within-document KG for each document based on an IE system (Lin et al., 2020), and then connect the within-document KGs into a cross-document KG using cross-document event coreference resolution. Based on the cross-document KG, we use a heterogeneous GNN (Schlichtkrull et al., 2018; Hu et al., 2019) to conduct detection. We further
incorporate the results of event-level detection to help the document-level detector.

4.1 Knowledge Graph Construction

Within-document KG: We first construct a within-document IE-based knowledge graph for each document. We leverage OneIE (Lin et al., 2020), a joint IE system, to extract the entities, relations and events contained in a given document. Then, we conduct entity linking and entity coreference resolution (Lee et al., 2017; Wen et al., 2021) to merge multiple mentions of the same entities together. Eventually, we obtain a within-document KG where entities and events are nodes, relations are edges between entities, and arguments are edges between events and entities.

Cross-document KG: We leverage cross-document event coreference resolution to connect the within-document KGs into a cross-document KG as illustrated in Figure 2. We employ a cross-document event coreference resolution system (Lai et al., 2021; Wen et al., 2021) to identify clusters of events from multiple documents that refer to the same real-world events. The system utilizes both textual contexts of the event mentions and the symbolic features such as the event type information. An example of the detected event cluster is shown in Table 1, where the four events from four documents all refer to the same explosion attack on President Nicolas Maduro. These four events contain complementary or contradictory details, which can be used for misinformation detection. For each event cluster, we add a node to represent the overall information of the real-world complex event corresponding to the cluster. Then, an edge is added between each event node and corresponding cluster node to allow reasoning among cross-document coreferential events.

To indicate which document each entity or event belongs to and capture the global information of each document, we further introduce a document node and connect it to the associated entity and event nodes for each document.

The resulting KG contains 4 types of nodes (i.e. entity nodes, event nodes, document nodes, and event cluster nodes) and 5 types of edges (i.e. relation edges, event argument edges, document-to-entity edges, document-to-event edges, and edges connecting event nodes to event cluster nodes). Since all edges are directional, we add an inverse edge for each edge to propagate features along both directions, and the final KG contains 10 edge types, accounting for the inverse of existing edge types.

Table 1: An example of cross-document event cluster from IED dataset, where trig, arg1 and arg2 represent the trigger, ExplosiveDevice argument and Place argument respectively. The four events from four documents all refer to the explosion attack on Nicolas Maduro. The two real news complement each other by providing different aspects of the news (ExplosiveDevice argument in the first news and Place argument in the second news), while the two fake news contradict the real news with different details (i.e., different ExplosiveDevice and Place arguments).

| Type   | Details |
|--------|---------|
| Real   | Venezuela’s president, Nicolás Maduro, has survived an apparent assassination attempt after what officials described as drones armed with explosives detonated overhead during a speech he was making at a military event. |
| Real   | The BBC quotes anonymous firefighters at the scene who say “the incident was actually a gas tank explosion inside an apartment,” but did not provide further details. |
| Fake   | Maduro was not targeted by the drones, the prime minister said, but state security services reported that the drones were meant for him. “The explosion was caused by two machine guns,” Maduro said, adding that there were no injuries. |
| Fake   | Two drones armed with explosives detonated near PuntoDeCorte, where the Venezuelan Foreign Minister, Jorge Rodríguez, was performing, and near the stage where he was giving a speech. |

KG representation: To initialize the node and edge embeddings in the KG, we use BERT (Devlin et al., 2019) to encode the text descriptions of nodes or edges and take the embeddings of [CLS] tokens. For a document node, we use the entire document as the text description; for an entity node, we use its canonical mention; for an event node, we use the sentence where the event trigger occurs; and for an event cluster node, we average the embeddings of all events in the cluster as the embedding for the cluster node. For a relation edge or an event argument role edge, we use the linearized representation as the text description. For example, the Leadership relation between “Nicolas Maduro” and “Venezuelan” is described as “Nicolas Maduro, Leadership, Venezuelan”, and “guns” as the ExplosiveDevice argument of the DetonateExplode event is described as “DetonateExplode, ExplosiveDevice, guns”.

4.2 Knowledge Graph Encoder

Heterogeneous GNN: Given the heterogeneous nature of the cross-document KG, we adopt a het-
ereogeneous GNN to encode the KG.

Formally, let $\mathcal{G}$ denote the KG and $\mathcal{V}$ denote the nodes in $\mathcal{G}$. We use $\mathcal{R}$ to denote the 10 types of edges as discussed in the previous section, and for each edge type $r \in \mathcal{R}$, we use $\mathcal{G}_r$ to denote the subgraph of $\mathcal{G}$ that only contains edges of type $r$. At the $l$-th layer, the inputs are output features produced by the previous layer denoted as $h^{(l-1)}_{\cdot,\cdot}$, $i \in \mathcal{V}$. For each edge type $r \in \mathcal{R}$, we apply a separate GNN to encode $\mathcal{G}_r$ and produce a set of features denoted as $h^{(l)}_{\cdot,\cdot,\cdot}$. Then, we aggregate the outputs for all edge types into the final output as follows:

$$h^{(l)}_{i,\cdot,\cdot} = \sum_{r \in \mathcal{R}} h^{(l)}_{i,\cdot,\cdot} / |\mathcal{R}| \quad (1)$$

For document-to-entity edges, document-to-event edges, and edges connecting event nodes to event cluster nodes, we use standard graph attention network (GAT). For relation edges and event argument edges, we apply edge-aware GAT to leverage the edge features. Here, the edge features refer to the BERT embeddings of text descriptions such as “Nicolas Maduro, Leadership, Venezuelan” or “DetonateExplode, ExplosiveDevice, guns” as described in Section 4.1. The remainder of Section 4.2 presents details of GAT and edge-aware GAT, i.e., how to produce $h^{(l)}_{i,\cdot,\cdot}$ based on $h^{(l-1)}_{i,\cdot,\cdot}$.

**Graph attention network:** For each given node, GAT aggregates the node features of its neighbors via attention mechanism (Velickovic et al., 2018). For a given edge type $r \in \mathcal{R}$, let $\mathcal{N}_{i,\cdot,\cdot}$ denote the neighbors of node $i$ in $\mathcal{G}_r$. At the $l$-th layer, the attention weights $\alpha_{ij}$ are calculated as follows:

$$e_{ij} = \text{LeakyReLU} \left( a^\top \left[ Wh^{(l-1)}_i \parallel Wh^{(l-1)}_j \right] \right) \quad (2)$$

$$\alpha_{ij} = \text{softmax}_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_{i,\cdot,\cdot}} \exp(e_{ik})} \quad (3)$$

where $a$ and $W$ are trainable parameters, and $\|$ denotes the feature concatenation. The output features $h^{(l)}_{i,\cdot,\cdot}$ for node $i$ in $\mathcal{G}_r$ are calculated as follows:

$$h^{(l)}_{i,\cdot,\cdot} = \sum_{j \in \mathcal{N}_{i,\cdot,\cdot}} \alpha_{ij} Wh^{(l-1)}_j \quad (4)$$

**Edge-aware graph attention network:** Edge-aware GAT is an extension of GAT that considers edge features in addition to node features (Huang et al., 2020; Yasunaga et al., 2021). Let $r_{ij}$ denote the features of the edge between node $i$ and $j$. For a given edge type $r \in \mathcal{R}$, at the $l$-th layer, the attention weights $\alpha_{ij}$ are computed as follows:

$$r'_{ij} = Wh \left[ h^{(l-1)}_{i,\cdot,\cdot} \parallel h^{(l-1)}_{j,\cdot,\cdot} \parallel r_{ij} \right] \quad (5)$$

$$\alpha_{ij} = \text{softmax}_j \left( (W^G h^{(l-1)}_i) (W^K r'_{ij})^\top \right) \quad (6)$$

where $W^V$, $W^K$, and $W^K$ are trainable parameters. The output features $h^{(l)}_{i,\cdot,\cdot}$ for node $i$ in $\mathcal{G}_r$ are computed as follows:

$$h^{(l)}_{i,\cdot,\cdot} = \sum_{j \in \mathcal{N}_{i,\cdot,\cdot}} \alpha_{ij} W^V r'_{ij} \quad (7)$$

where $W^V$ is a learnable matrix.

### 4.3 Misinformation Detector

Using the previously described graph encoder, we are able to obtain representations of the document and event nodes. We conduct document-level detection using the document node representations, and event-level detection using the event node representations. We separately train two detectors for these two levels of tasks.

However, these two tasks are not mutually independent. Intuitively, document-level detection can benefit from the results of event-level detection, because the presence of a large number of false events indicates that the document is more likely to be fake. Therefore, we feed the results produced by a well-trained event-level detector into each layer of the document-level detector. Let $e_i$ denote the representations of node $i$ produced by the event-level detector. At the $l$-th layer of the document-level detector, instead of using the output features of the previous layer $h^{(l-1)}_{i,\cdot,\cdot}$ as input features, we use a linear projection of the concatenation of $e_i$ and $h^{(l-1)}_{i,\cdot,\cdot}$ calculated as follows:

$$W^\text{proj}_{i} \left[ e_i \parallel h^{(l-1)}_{i,\cdot,\cdot} \right] \quad (8)$$

where $W^\text{proj}_{i}$ is a learnable matrix.

### 5 Dataset Construction

Currently, there are no existing resources for cross-document misinformation detection. We propose to construct datasets based on real news datasets with clustering information. For each cluster, we randomly sample 50% real news and replace them with manipulated fake news. Figure 3 shows an...
We construct 3 new benchmark datasets based on (1) we only conduct entity swapping, and do not
therefore, a complex event can be considered as a
document cluster. The main differences between Fung et al. (2021)’s
described by multiple documents (Li et al., 2021).

cally related documents.

a real-world story (e.g., Boston bombing) and is
described by multiple documents (Li et al., 2021).
Therefore, a complex event can be considered as a
document cluster. TL17 and Crisis are two timeline
summarization datasets containing multiple
“timelines”. Each timeline contains multiple
documents describing an evolving long-term event such as Influenza H1N1 and Egypt Revolution (Tran et al., 2013, 2015), and thus can be regarded as a
document cluster. The detailed statistics of the
original datasets are shown in Appendix A.

However, the documents within the same cluster
can not be closely related as the story described
may not span up to three years. To obtain
smaller and more closely related clusters, we split
each timeline into smaller clusters of approximately
size of 10 based on publication dates. Then, we
employ the methods described in Section 5 to gen-
erate fake documents. The statistics of the con-
structed datasets are in Table 2.

6.2 Experimental Settings

For our proposed method, we use a 4-layer hetero-
geneous GAT and use bert-base-uncased to
initialize the node and edge embeddings. For
comparison, on the document-level detection task,
we compare our method against two baselines:
HDSF that models inter-sentence dependency tree
(Karimi and Tang, 2019), and GROVER (Zellers et al., 2019), a Transformer-based detector. On
the event-level detection task, since there are no exist-
ings methods, we compare our method against two
heuristic baselines: random guessing and logistic
regression. For logistic regression, we use hand-
crafted features to represent the event including
the event type, the number of arguments, and the
size of the event cluster. The detailed settings are
presented in Appendix B.

For evaluation, we use F1 to evaluate document-
level detection. Considering the label imbalance
of event-level detection, we use F1 and AUC to
evaluate event-level detection. For F1 metric, we

Table 2: Statistics of the resulted datasets.

| Dataset | # Cluster | # Doc | # Fake event per doc (%) |
|---------|-----------|------|-------------------------|
| IED     | Train     | 422  | 3.99 (9.91%) |
|         | Dev       | 140  | 3.66 (9.14%) |
|         | Test      | 140  | 3.68 (9.51%) |
| TL17    | Train     | 276  | 2.97 (12.70%) |
|         | Dev       | 92   | 2.69 (12.31%) |
|         | Test      | 92   | 2.85 (12.13%) |
| Crisis  | Train     | 1413 | 4.54 (13.95%) |
|         | Dev       | 177  | 4.21 (13.29%) |
|         | Test      | 177  | 4.38 (13.80%) |

For IED, we randomly split the clusters due to the lack of
publication dates.

6 Experiments

6.1 Data

We construct 3 new benchmark datasets based on
three datasets that naturally have clusters of topi-
cally related documents. IED dataset is a complex
event corpus, where each complex event refers to
a real-world story (e.g., Boston bombing) and is
described by multiple documents (Li et al., 2021).
Therefore, a complex event can be considered as a
document cluster. TL17 and Crisis are two timeline
summarization datasets containing multiple

... The device exploded over an area where uniformed guardsmen stood at attention in ranks, injuring seven. Advertisement ...

KG of fake news

DetonateExplode

IED

DetonateExplode

device

machine guns

Fake device

Fake device
We report the F1 scores of HDSF (Karimi and Tang, 2019), GROVER of two settings (Zellers et al., 2019), and our proposed method.

| Cross-document event coreference | Event-level detection results | IED | TL17 | Crisis |
|----------------------------------|-------------------------------|-----|------|--------|
|                                  |                               | 80.59 | 86.55 | 93.64 |
| ✗                                | ✗                             | 84.57 | 88.99 | 93.67 |
| ✗                                | Random                        | 83.63 | 84.86 | 92.18 |
| ✗                                | ✗                             | 86.76 | 90.21 | 93.89 |

Table 3: F1 results (in %) of document-level detection. We report the F1 scores of HDSF (Karimi and Tang, 2019), GROVER of two settings (Zellers et al., 2019), and our proposed method.

Table 4: F1 results (in %) of ablation study over document-level detection. We analyze the use of cross-document event coreference resolution and event-level detection results. We further experiment with random features for event-level detection results. Results of our full method are presented in the last row.

select the optimal threshold on the validation set.

6.3 Document-level Detection Results

Table 3 shows the results of document-level detection. Our method yields consistent improvements on all 3 datasets, and significantly outperforms the baselines that judge the authenticity for each document in isolation. To understand the effectiveness of each component, we conduct an ablation study and show the results in Table 4. We have the following findings:

1. We remove the edges between event nodes and event center nodes to analyze the impact of cross-document event coreference resolution, and find that such information significantly improves the performance on IED and TL17. We also train our detector with smaller clusters on TL17 and get worse performance (84.53% and 87.37% on clusters with size 1 and 2 respectively), which verifies that our model benefits from more cross-document information. The benefit of cross-document event coreference resolution is less significant on the large-scale Crisis dataset containing 1.7k documents. This may imply that cross-document misinformation detection is more useful for emerging new events where large-scale training data is not available.

2. Using the event-level detection results consistently improves the performance by 1-3 points on all datasets. Since the projection modules introduce additional parameters, we further train a detector utilizing random features and find that using random features reduces the performance. This verifies that the improvement is brought by utilizing the knowledge learnt by the event-level detector rather than additional parameters.

6.4 Event-level Detection Results

We track the manipulation operations during the dataset construction process, which allows us to obtain supervision for event-level detection. The results are shown in Table 5. Since there are no existing methods for this new task, we compare our method with two heuristic baselines, random guessing and logistic regression with hand-crafted event features. We find that random guessing performs the worst, logistic regression achieves satisfactory performance, and our method significantly outperforms these two baselines by a large margin. As in document-level detection, we conduct an ablation study on the use of cross-document event coreference resolution by removing edges between event nodes and event cluster nodes, and find that such information brings slight improvements over AUC metric.

6.5 Analysis and Discussion

To demonstrate the benefits of using cross-document event coreference resolution, we show an example in Figure 4, with 4 documents from the same cluster. As shown in Figure 4, by performing cross-document reasoning on events in the same event cluster, our model achieves better performance compared to Ours(ABLATION), i.e., our model without edges between event nodes and event cluster nodes.

We further analyze the remaining errors from our model. Figure 5 shows two representative cases
Figure 4: An example of four documents from the same cluster in the IED dataset. Event triggers are bolded and marked in gold, and fake information is marked in red. The tables report the detection results of our model with and without cross-document event coreference resolution, denoted by “Ours” and “Ours (ABLATION)” respectively, and better results are bolded. The use of cross-document event coreference resolution significantly enhances both levels of detection, especially for detecting fake news 1.

Figure 5: Two examples where our detector fails to detect the fake information. Event triggers are bolded and marked in gold, and fake information is marked in red. In the first example, the fake event argument Abqaiq City is not captured by the IE system and thus cannot be detected. In the second example, the visit of Vajpayee to Mumbai is fake information but not mentioned by any other documents, and no coreference is detected for the Transportation event. Therefore, our detector does not have enough information to detect the fake information.

where both document-level and event-level detectors fail to detect misinformation. In the first example, the manipulated entity is not captured by the IE system, and the error of IE system is propagated into the detector. A potential solution is to use an OpenIE system (Stanovsky et al., 2018) that is able to cover more event and entity types. The second example is a more challenging case where the event containing fake information is not mentioned by any other documents. This makes it difficult to either verify or disprove via cross-document reasoning, and may require the detector to actively search for external information related to the event.

There are some remaining challenges and limitations in our proposed methodology. First, some cross-document contradictions are difficult to capture by coreference resolution only. In the example in Figure 1, knowing that the police is unlikely to help and attack Boyland at the same time requires commonsense knowledge reasoning, which we leave as our future work. Second, an underlying assumption of our framework is that real news articles are consistent and complementary with each other, while fake news often contradicts each other. This assumption is true for our constructed datasets because we manipulate the KGs via random entity swapping. However, certain types of human-written fake news documents, such as conspiracy theories, tend to be closely related to each other and convey highly similar information because they share the same biases or aim to manipulate readers in the same way. This may limit the performance of our proposed system in real-world scenarios.

7 Conclusions and Future Work

We are the first to study the new task of cross-document misinformation detection. We conduct the task at two levels, document-level and the more fine-grained event-level, and construct 3 new datasets to handle the lack of training data. We further propose a graph-based cross-document detector that conducts reasoning over a cross-document knowledge graph and feed the event-level detection results into document-level detector. Experimental results show that our proposed method significantly outperforms existing methods.

For future work, we intend to extend our method to conduct cross-document reasoning over more types of information (e.g., entities and relations) in addition to events. We also plan to extend our method to multi-media news including texts, images, audios, and videos, which requires the construction of cross-document multi-modal knowledge graphs. Finally, a challenging but important task is to construct a large-scale fake news detection corpus with human-written fake news containing document clusters and study our method in this scenario.
8 Ethical Considerations

The goal of this work is to advance state-of-the-art research in the field of misinformation detection, by analyzing multiple documents on the same topic. We build new benchmark datasets using a fake news generator, and propose a detector that achieves high performance in such scenarios. We have released the constructed datasets and detector codes in this submission as a useful reference for future research. We hope our work will encourage more efforts in this direction and benefit the community.

However, as with any work that utilizes text generation, our work involves the risk of being applied to produce false information to mislead or manipulate readers. Therefore, we promise not to share codes or checkpoints of our generator to avoid potential negative consequences. To improve reproducibility, we describe the general idea and a few crucial details of the fake news generator.

References

Pepa Atanasova, Preslav Nakov, Lluís Márquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James R. Glass. 2019. Automatic fact-checking using context and discourse information. ACM J. Data Inf. Qual., 11(3):12:1–12:27.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics.

Thomas Diggelmann, Jordan L. Boyd-Graber, Janis Bullan, Massimiliano Ciaramita, and Markus Leipold. 2020. CLIMATE-FEVER: A dataset for verification of real-world climate claims. CoRR, abs/2012.00614.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen R. McKeown, Mohit Bansal, and Avi Sil. 2021. Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pages 1683–1698. Association for Computational Linguistics.

Linmei Hu, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. 2019. Heterogeneous graph attention networks for semi-supervised short text classification. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 4820–4829. Association for Computational Linguistics.

Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021. Compare to the knowledge: Graph neural fake news detection with external knowledge. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pages 754–763. Association for Computational Linguistics.

Kung-Hsiao Huang, Mu Yang, and Nanyun Peng. 2020. Biomedical event extraction on graph edge-conditioned attention networks with hierarchical knowledge graphs. In Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020, volume EMNLP 2020 of Findings of ACL, pages 1277–1285. Association for Computational Linguistics.

Hamid Karimi, Proteek Roy, Sari Saba-Sadiya, and Jiliang Tang. 2018. Multi-source multi-class fake news detection. In Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018, pages 1546–1557. Association for Computational Linguistics.

Hamid Karimi and Jiliang Tang. 2019. Learning hierarchical discourse-level structure for fake news detection. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 3432–3442. Association for Computational Linguistics.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: multimodal variational autoencoder for fake news detection. In The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019, pages 2915–2921. ACM.

Tuan Manh Lai, Heng Ji, Trung Bui, Quan Huang Tran, Franck Dernoncourt, and Walter Chang. 2021. A context-dependent gated module for incorporating symbolic semantics into event coreference resolution. In Proceedings of the 2021 Conference of the
Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. Automatic detection of fake news. In Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018, pages 3391–3401. Association for Computational Linguistics.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, pages 188–197. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghashvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 7871–7880. Association for Computational Linguistics.

Manling Li, Sha Li, Zhenhailong Wang, Lifu Huang, Kyunghyun Cho, Heng Ji, Jiawei Han, and Clare R. Voss. 2021. Future is not one-dimensional: Graph modeling based complex event schema induction for event prediction. CoRR, abs/2104.06344.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 7999–8009. Association for Computational Linguistics.

Jamal Abdul Nasir, Osama Subhani Khan, and Iraklis Varlamis. 2021. Fake news detection: A hybrid cnm-rnn based deep learning approach. International Journal of Information Management Data Insights, 1(1):100007.

Van-Hoang Nguyen, Kazunori Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. FANG: leveraging social context for fake news detection using graph representation. In CIKM ’20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020, pages 1165–1174. ACM.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In Proceedings of NAACL-HLT 2019: Demonstrations.

Jeff Z. Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu. 2018. Content based fake news detection using knowledge graphs. In The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, Monterey, CA, USA, October 8-12, 2018, Proceedings, Part I, volume 11136 of Lecture Notes in Computer Science, pages 669–683. Springer.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018, pages 3391–3401. Association for Computational Linguistics.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, pages 2931–2937. Association for Computational Linguistics.

Victoria L. Rubin, Niall Conroy, Yinmin Chen, and Sarah Cornell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In Proceedings of the second workshop on computational approaches to deception detection, pages 7–17.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. Covid-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pages 2116–2129. Association for Computational Linguistics.

Sohan De Sarkar, Fan Yang, and Arjun Mukherjee. 2018. Attending sentences to detect satirical fake news. In Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018, pages 3371–3380. Association for Computational Linguistics.

Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Wellin. 2018. Modeling relational data with graph convolutional networks. In The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings, volume 10843 of Lecture Notes in Computer Science, pages 593–607. Springer.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data, 8(3):171–188.

Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019, pages 312–320. ACM.
Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pages 885–895. Association for Computational Linguistics.

Reuben Tan, Bryan A. Plummer, and Kate Saenko. 2020. Detecting cross-modal inconsistency to defend against neural fake news. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, pages 2081–2106. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pages 809–819. Association for Computational Linguistics.

Giang Binh Tran, Mohammad Alrifai, and Eelco Herder. 2015. Timeline summarization from relevant headlines. In Advances in Information Retrieval - 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015. Proceedings, volume 9022 of Lecture Notes in Computer Science, pages 245–256.

Giang Binh Tran, Mohammad Alrifai, and Dat Quoc Nguyen. 2013. Predicting relevant news events for timeline summaries. In 22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013. Companion Volume, pages 91–92. International World Wide Web Conferences Steering Committee / ACM.

Vaibhav Vaibhav, Raghuram Mandyam Annasamy, and Eduard H. Hovy. 2019. Do sentence interactions matter? leveraging sentence level representations for fake news classification. In Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing, TextGraphs@EMNLP 2019, Hong Kong, November 4, 2019, pages 134–139. Association for Computational Linguistics.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018. Conference Track Proceedings. OpenReview.net.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, pages 7534–7550. Association for Computational Linguistics.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers, pages 422–426. Association for Computational Linguistics.

Haoyang Wen, Ying Lin, Tuan Lai, Xiaoman Pan, Sha Li, Xu Dong Lin, Ben Zhou, Manling Li, Haoyu Wang, Hongming Zhang, Xiaodong Yu, Alexander Dong, Zhenhai Long Wang, Yi Fung, Piyush Mishra, Qing Lu, Didac Surís, Brian Chen, Susan Windisch Brown, Martha Palmer, Chris Callison-Burch, Carl Vondrick, Jiawei Han, Dan Roth, Shih-Fu Chang, and Heng Ji. 2021. RESIN: A dockerized schema-guided cross-document cross-lingual cross-media information extraction and event tracking system. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations, NAACL-HLT 2021, Online, June 6-11, 2021, pages 133–143. Association for Computational Linguistics.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: reasoning with language models and knowledge graphs for question answering. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, pages 535–546. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 9051–9062.

A Statistics of Original Datasets

Statistics of the original IED, TL17 and Crisis dataset are presented in Table 6.

B Experiment Details

Detailed settings of our method: For our proposed method, we use a 4-layer heterogeneous GNN, where each GAT layer contains 8 heads. To initialize the node and edge embeddings, we use bert-base-uncased model with the feature...
Table 6: Statistics of the original datasets.

|    | # Cluster | # Doc | # Doc per cluster |
|----|-----------|------|-------------------|
| IED | 433       | 7403 | 17                |
| TL17| 17        | 4650 | 273               |
| Crisis | 4   | 20463| 5116              |

dimension of 768. Our model contains 233M parameters.

For hyperparameters, we use a batch size of 16, and search the learning rate from \{10^{-3}, 10^{-4}, 10^{-5}\} and the number of layers within \{2, 4, 8\}. Our best-found hyperparameters are a learning rate of \(10^{-5}\) and a number of layers of 4. We train our model with Adam optimizer until convergence. To reduce computation cost, we freeze BERT’s parameters. The training process takes approximately 6 hours on a Tesla P100 GPU.

**Document-level baselines:** For document-level detection, we compare our method against two baselines: HDSF that models inter-sentence dependency tree (Karimi and Tang, 2019), and GROVER (Zellers et al., 2019), a Transformer-based detector. For HDSF, we use the implementation at https://github.com/hamidkarimi/HDSF/. For GROVER, we use the implementation at https://github.com/rowanz/grover and experiment with two settings, medium setting and mega setting. Since fine-tuning the GROVER model is computationally expensive, we use GROVER in the zero-shot setting.

**Event-level baselines:** For event-level detection, since there are no existing methods, we use two heuristic baselines, random guessing and logistic regression. In random guessing, for each event, we randomly draws a value from a uniform distribution between [0, 1] as the probability that the event is false. In logistic regression, we use the following features: event type (represented by one-hot feature), number of arguments, and number of coreferential events. The features are normalized on the training set. We use the implementation of logistic regression and default parameters provided by sklearn.

**C Examples of Fake News Generation**

We present two examples of generated fake news in Figure 6 and 7, including the original real news, manipulated KG, and generated fake news. The generated fake news conveys the manipulated misinformation and meanwhile is stylistically similar to real news.

**D Scientific Artifacts**

In this work, we use three datasets including IED (Li et al., 2021), TL17 (Tran et al., 2013) and Crisis (Tran et al., 2015). There are no licenses or terms of use associated with all three datasets.

We use five software. Among them, HDSF (Karimi and Tang, 2019), OneIE (Lin et al., 2020) and RESIN (Wen et al., 2021) have no license or terms of use. GROVER (Zellers et al., 2019) and huggingface are licensed under the Apache License 2.0. Fairseq (Ott et al., 2019) is licenced under the MIT License.

We use two models, BERT (Devlin et al., 2019) and BART (Lewis et al., 2020), licenced under the Apache License 2.0 and the MIT License respectively.

In summary, all artifacts involved either have no associated licenses or terms of use, or are licensed under the Apache License 2.0 or the MIT License. Both the Apache License 2.0 or the MIT License permit commercial and private use. Therefore, our use is consistent with their intended use. We will release the dataset and software with licenses compatible with the original access conditions.
Figure 6: An example of generated fake news, including the original real news, manipulated KG, and generated fake news. Real and fake information are marked in blue and red respectively. To save space, we only show some parts of the KG that are manipulated.
Figure 7: An example of generated fake news, including the original real news, manipulated KG, and generated fake news. Real and fake information are marked in blue and red respectively. To save space, we only show some parts of the KG that are manipulated.