InfoSurgeon: Cross-Media Fine-grained Information Consistency Checking for Fake News Detection

Yi R. Fung[1], Chris Thomas[2], Revanth Reddy[1], Sandeep Polisetty[3], Heng Ji[1], Shih-Fu Chang[2], Kathleen McKeown[2], Mohit Bansal[4], Avirup Sil[5]
[1]University of Illinois at Urbana-Champaign, [2]Columbia University
[3]UMass Amherst, [4]University of North Carolina at Chapel Hill, [5]IBM
[1]{yifung2,revanth3,hengji}@illinois.edu
[2]{christopher.thomas,sc250,kathy}@columbia.edu
[3]spolisetty@umass.edu, [4]mbansal@cs.unc.edu, [5]avi@us.ibm.com

Abstract

To defend against neural system-generated fake news, an effective mechanism is urgently needed. We contribute a novel benchmark for fake news detection at the knowledge element level, as well as a solution for this task which incorporates cross-media consistency checking to detect the fine-grained knowledge elements making news articles misinformative. Due to training data scarcity, we also formulate a novel data synthesis method by manipulating knowledge elements within the knowledge graph to generate noisy training data with specific, hard to detect, known inconsistencies. Our detection approach outperforms the state-of-the-art (up to 16.8% absolute accuracy gain), and more critically, yields fine-grained explanations.[1]

1 Introduction

In recent years, generative neural network models in natural language processing (Zellers et al., 2019) and computer vision (Choi et al., 2018) have become the frontier for malicious actors to controllably generate misinformation at scale. These realistic-looking AI-generated "fake news" have been shown to easily deceive humans, and it is, thus, critical for us to develop robust verification techniques against machine-generated fake news (Tan et al., 2020; Zellers et al., 2019; Kaliyar et al., 2020). Current misinformation detection approaches mainly focus on document-level fake news detection using lexical features and semantic embedding representations (Wang, 2017; Karimi et al., 2018; Tan et al., 2020). However, fake news is often generated based on manipulating (misusing, exaggerating, or falsifying) only a small part of the true information, namely the knowledge elements (KEs, including entities, relations and events). Moreover, recent news oftentimes makes claims that do not have verified evidence yet, and evaluating the truthfulness of these real-time claims depends more on their consistency with other information conveyed in other data modalities.

In this paper, we propose a new task: *fine-grained, knowledge element-level cross-media information consistency checking*. The task involves treating the entire multimedia news article as one whole interconnected claim, where the goal is to detect misinformative KEs across the image, caption, and body text, as revealed by inconsistencies with respect to itself, or to background knowledge. This KE-level detection approach directly points out the fake pieces of information in the news, allowing for better explainability.

Figure 1 shows an example where both the text and image provide complementary information about key argument roles of an event. We present the **Info**rmation **Surgeon** **(InfoSurgeon)** model, which takes full advantage of state-of-the-art multimedia joint knowledge extraction techniques to analyze fine-grained event, entity, and relation elements, as well as whether these extracted KEs align consistently across modalities and background knowledge. We propose a novel probabilistic graphical neural network model to fuse the outputs from these indicators.

A major challenge to performing KE level misinformation detection is the lack of training data. Hence, we additionally propose a novel approach to generate noisy training data automatically since existing fake news generators (Zellers et al., 2019) do not track the specific pieces of information generated that are fake. We take a real news article, extract a multimedia knowledge graph, and replace or insert salient nodes or edges in the graph. We track the specific manipulation operations, and regenerate the manipulated version of the news article

[1]The code, data and resources related to the misinformation detector are made publicly available at https://github.com/yrf1/InfoSurgeon for research purposes.

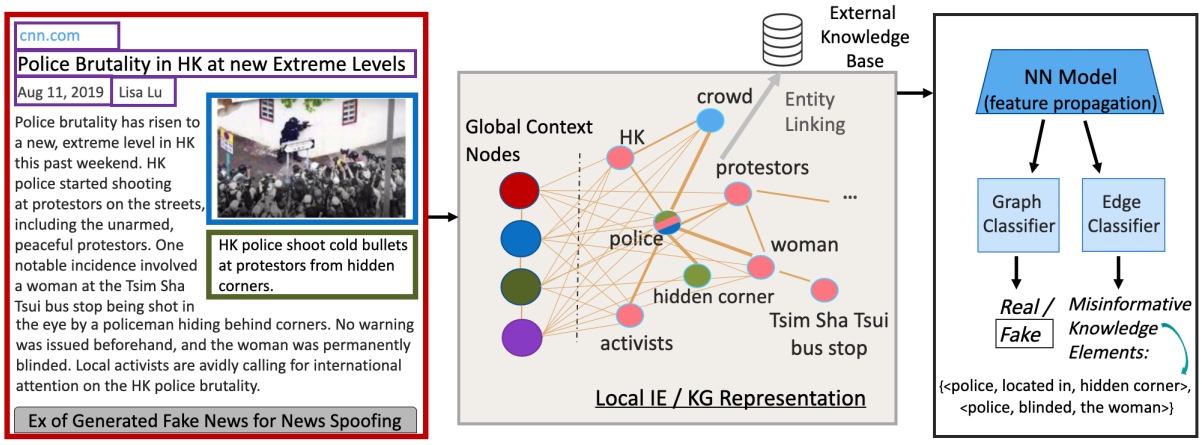

Figure 1: The architecture of our fine-grained, KE-level fake news detection system, *InfoSurgeon*. Our model uses graph-based neural network to aggregate cross-modal and external knowledge in a multimedia KG to determine whether a document is real or fake and provide KE-level explanations. For instance, the document above should be detected as fake due to cross-modal inconsistencies (i.e. the caption conveys *<police, located.in, hidden corner>* to surprise-attack protesters, which is inconsistent with the image showing *<police, located.near, visible crowd of reporters>*). *<police, blinded, woman>* in the article is also fake information, which is not supported by the image or caption. Note: the article in this figure is inspired from materials reported in major news outlet that were later taken down due to misinformation.

using a graph-to-text approach (Ribeiro et al.,2020), while filtering out poor quality unconvincing generations through a neural adversarial filter.

Experiment results show that our approach achieves 92%-95% detection accuracy, 16.8% absolute higher than the state-of-the-art approach (Tan et al., 2020). Ablation tests demonstrate the effectiveness of our new detection method. The major contributions of this paper are:

- We propose a novel approach to perform fake news detection at the KE level, representing the claims in the news article as a multimedia knowledge graph and detecting the mis-informative pieces in the form of KEs for a strong explainability.
- We contribute ***InfoSurgeon***, a unified framework for detecting misinformation in news articles that comprehensively incorporates source context, semantic representation, multimedia information elements, and background knowledge in a reasoning framework.
- Finally, we present a novel benchmark, KE level fake news detection, with a silver standard annotation dataset (15,000 multimedia document pairs) automatically generated by KG conditioned natural language generation.

2 Task Formulation

Given a multimedia news article, X, which consists of its body text bt, list of images $im_{1..i}$, list of accompanying captions $c_{1..i}$, and metadata $m = (domain, date, author, headline)$, our study aims to detect the presence of misinformation at two levels. In document-level detection, we classify each news article as either *real* or *fake*, overall. In knowledge element-level detection, we predict the specific set of knowledge elements in the news article conveying misinformation. Here, we refer to *knowledge elements* (KEs) comprehensively as the entities, relations, events, and subgraphs/metapaths (Fu et al., 2020) in an information network.

To detect the misinformative KEs, we treat each news article as one ultimate claim represented by a multimedia knowledge graph $KG = (N, E)$ capturing the important information conveyed. The nodes (N) in the KG consist of entities (t), while the edges (E) in the KG consist of relations (r) or event argument roles (a) connecting the entities. Detecting the KEs causing a news article to be fake boils down to extracting *<subject_entity, predicate, object_entity>* triplets from the multimedia input data, and labeling all of the triplets in which the relation or event between a head entity and tail entity holds false as evidence of misinformation through binary edge classification. Figure 2 shows exam-

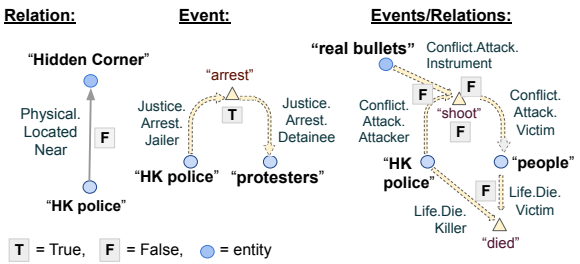

Figure 2: In the case of events, we ignore the event trigger denoted by △ and connect entities by their event argument roles and event types combined *e.g.*, *<HK police, Justice.Arrest.Jailer-Justice.Arrest.Detainee, (uncooperative) protesters>*. The True/False tags are labeled for each triplet, which connects a pair of entities.

ples of how KEs should be detected if they occur in the KG of a news article.

We evaluate document-level fake news detection based on the established metric of prediction accuracy. To evaluate fine-grained KE level fake news detection, we compute the F-score: the harmonic mean of precision and recall across KEs. This is an appropriate metric due to the imbalanced nature of fake KEs, which usually constitute the minority.

3 Fake News Detection

3.1 Overview

As shown in Figure 1, our fine-grained multimedia fake news detection system, *InfoSurgeon*, extracts features from both global context and local KG. The global context nodes capture the semantic representations of the body text, images, captions, and metadata in the news article. The local KG provides an explicit representation of the key information pieces and their interactions. As a clarifying example, the entire image in a news article constitutes a global context node, while the specific objects detected in the image make up the node entities in the local KG. *InfoSurgeon* combines these two complementary components by connecting the global context nodes to the entity nodes in the KG extracted from the news article, thereby propagating context signals into the knowledge elements.

3.2 Global Context Representation

To incorporate general context information and take advantage of cross-media inconsistencies that are more likely to exist in fake news, we compute semantic representations for each news article component to initialize the node features. We feed the body text and each caption through the summarization-based BERT encoder from Liu and Lapata (2019), which averages the encoded token embeddings across sentences through a weighted mechanism. For metadata, we run the text encoder on a string containing the article domain, publication date, author, and headline. For images, we concatenate object-based (Anderson et al., 2018) and event-based (Pratt et al., 2020) visual features. Features for the edges between global context nodes are initialized by the attention-based semantic similarity between the node features (Tan et al., 2020).

3.3 Local KG Representation

Constructing a KG from each Multimedia News Article: We leverage a publicly available multimedia Information Extraction (IE) system (Li et al., 2020; Lin et al., 2020) to construct a within-document knowledge graph $KG = (N_t, E_{r|a})$ for each multimedia article. The IE system can extract 197 types of entities, 61 types of relations, and 144 types of events from text and images. Then, it performs entity linking (Pan et al., 2017) to map entities extracted from both text and images to a background knowledge base e.g. Freebase (Bollacker et al., 2008) and NIL (unlinkable) entity clustering for name mentions that cannot be linked, followed by cross-media event and entity coreference resolution and grounding (Lai et al., 2021; Wen et al., 2021; Pratt et al., 2020).

Initializing the KG Embeddings: We define an attribute function, $\mathbf{A} : N_t, E_{r|a} \Rightarrow F$, that transforms each of the nodes and edges to its initial representation by concatenating the following features:

- Background Embeddings - For the entity nodes N_t that can be linked to Freebase, we use data dump from Google Developers resources[2] to map them to their respective Wikipedia pages, which serve as a rich source of established background knowledge. Background node embedding features are initialized from passing a Long Short Term Memory networks (LSTM) based architecture (Gers et al., 2000) through the word embeddings (Pennington et al., 2014) of the first paragraph in the Wikipedia page, which usually starts with a mention of the Wiki page's title. Background edge embedding features are initialized from passing the LSTM through the paragraphs that contain the mentions of both the head and tail nodes. These embeddings are set to a default zero vector for unlinkable nodes.

[2]https://developers.google.com/freebase/

- News Embeddings - These are the surface-level features circumstantial to the entities, relations, and events extracted. News-based node features are initialized from passing an LSTM through the word embeddings of the canonical entity mention extracted. News-based edge features are initialized from passing the LSTM through the word embeddings of the relation type or event argument role infused with head and tail entity information, in the triplet format e.g. "*<HK police, Physical.LocatedNear, visible crowd>*".
- Source Attribution - This is a 4-dimensional binary vector indicating whether the KE came from the body text, image, caption, or metadata.

3.4 Feature Propagation and Joint Learning

A central idea to our misinformation detector is that edge embeddings are naturally more closely aligned to the node embeddings they are connected to for the non-fake triplets. Therefore, we learn a neural network layer to extract the hidden representations of credibility between node connections. The graphical representation of the global context and local KG network is heterogeneous in nature though, so we propagate features as follows.

For the global context subgraph, potential misinformation lies in whether the images, captions, or metadata align with the overall news article. Given a global context node, u, we compute the hidden representations of credibility with all other global context node neighbors $v \in nbr(u)$ (1), and aggregate the information back to node u itself (2).

$$h_{e_{uv}} = \text{relu}\left(W_t \cdot [h_{n_u}, h_{e_{uv}}, h_{n_v}]\right) \quad (1)$$

$$h_{n_u} = \text{relu}\left(\frac{1}{|\text{nbr}(u)|} \sum_{v \in \text{nbr}(u)} h_{e_{uv}}\right) \quad (2)$$

For the local KG, potential misinformation lies in the relations or event argument roles connecting entity nodes. Given two local KG nodes u and v that are connected by an edge, we compute the hidden representation of triplet credibility as in eq (1). To further take advantage of neighborhood information, we propagate features across the global context and local KG network with graph attention and message passing.

3.5 Detector Component

Document Level Fake News Detection: An established approach to graph-level classification is to merge the extracted graph features together through AVG or MAX pooling. To strengthen signals, we further add primitive indicator values before the document level linear classifier. Tan et al. (2020) use a single binary indicator for the existence of overlap between entities in the caption and entities in the article body. We use a broader set of indicators reflecting the number of overlapping entities and events across the caption, body, and image.

Knowledge Element Level Fake News Detection: Detecting misinformative knowledge elements in the KG can be treated as a binary edge classification problem, in which each edge represents the entire triplet in which it serves as the predicate. We run a linear classifier on each of the learned edge embeddings that are not directly connected to the semantic nodes, to detect if the relation or event argument role connecting two entities is normal or not.

4 Fake News Generation

Currently, there exists no annotated dataset for KE level misinformation detection. A primary reason may be due to explicitly fake (as opposed to subtly biased) news being edited or taken down by online platforms after initial posting. Because manually labeling the misinformative KEs in a real-world news corpus is expensive, we aim to create a novel dataset with controlled synthesis of news articles and automatically generated labels for fine-grained KE level explainability in fake news detection. In this section, we propose two novel approaches that generate fake news, and at the same time, automatically label the misinformative knowledge elements. Given a set of real news articles, X_{real}, we perform deliberate edit operations on certain salient KEs in the new articles' KG to derive a manipulated representation, KG'. Hence, we can generate a new article conditioned on KG'. The corresponding KE level label can then be automatically derived, with the manipulated elements as fake and the unaltered elements as real, while the document-level label for the new generated article is fake.

4.1 Manipulated KG-to-text Synthesis

Given a pristine, real news article, we aim to perform controlled fake new synthesis by altering certain entities, relations, and events, while keeping the rest of the story largely intact. We observe that, in general, the entity nodes with the strongest degree of connection are the centerpiece of a news

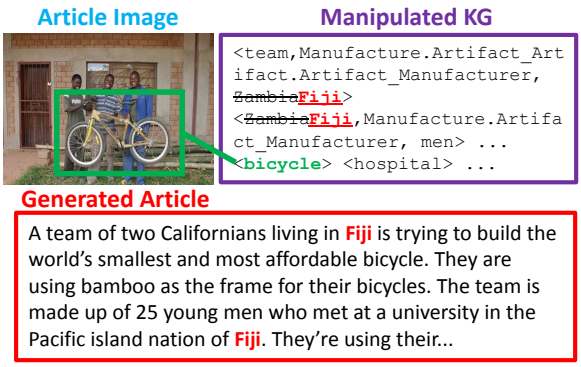

Figure 3: We show example of manipulating the multimedia KG of a news article, swapping geolocation-typed entity "Zambia" with "Fiji".

article, while the entity nodes with the smallest degree of connection are less salient. Thus, we randomly select entity nodes occurring at mid range frequency to manipulate. We vary the type of KG manipulation, as follows: (1) **Entity swapping** - we swap the original entity with an alternative entity that belongs to the same entity type. (2) **Addition of a new relation or event** - we take an existing entity, randomly select a relation or event argument role that connects to this entity type, and append a new entity at the other end of the relation or event. (3) **Subgraph replacement** - we select a subgraph of the news article that branches off the randomly selected entity nodes above, and replace it with a subgraph from another news article. Although we also considered the removal of node and edges, we intuitively found it too challenging to detect because lack of information can exist at various points across the article in reality but the silver standard annotation from selective removal would not cover enough of these for supervised training.

Next, we generate a fake news article that aligns with this manipulated KG' by finetuning a BART-large language model (Lewis et al., 2020) on our training set. To better enforce that manipulated entities actually appear in the generated article, we use a copy mechanism which re-purposes entities from the input KG when generating the output article (Post and Vilar, 2018). After training, we manipulate KGs as described above and feed the manipulated KGs into our model to generate synthetic data (see the example in Figure 3). Importantly, the manipulated knowledge elements serve as silver-standard annotations for the generated fake news articles.

4.2 Manipulated AMR-to-Text Synthesis

Tan et al. (2020) observe captions to be very significant in detecting fake articles, with performance dropping from 85.6% (when trained on articles generated using GROVER-Mega (Zellers et al., 2019)) to 56.9% when captions are excluded. Hence, we aim to further manipulate existing captions by generating subtle variations in the relations between entities. We leverage Abstract Meaning Representation (AMR) (Banarescu et al., 2013) graphs extracted from these captions since they capture rich fine-grained sentence-level semantic relations expressing *who does what to whom*. AMR semantic representation includes PropBank (Palmer et al., 2005) frames, non-core semantic roles, coreference, entity typing and linking, modality, and negation.

True Caption:
In Afghanistan, the Taliban released to the media this picture, which it said shows the suicide bombers who attacked the army base in Mazar-i-Sharif, April 21, 2017

Fake caption:
On 21 April 2017 the Taliban released this picture to the army in Afghanistan which they said was a suicide bomber hiding at a media base in the city of Mazar-i-Sharif

Figure 4: Example of AMR-to-text fake caption generation. The roles of *army* and *media* (in blue) are switched and the node corresponding to the event trigger (in red) *attacked* is negated.

To obtain the AMR graphs, we use the stack-transformer based AMR parser from Astudillo et al. (2020) and train it on AMR 3.0[3]. Given the AMR graph, we vary the manipulation as follows: (1) **Role switching** - we randomly select two entity mentions that are present in different argument subgraphs of the AMR root node and interchange their positions in the AMR graph. (2) **Predicate negation** - we randomly pick predicates in the AMR graph corresponding to event triggers and other verbs, and replace them with their antonyms, which we obtain from WordNet (Fellbaum, 1998). This manipulation also includes reverting nodes with negative polarity, thereby negating the sentence.

[3]https://catalog.ldc.upenn.edu/LDC2020T02

After manipulating the AMR graphs, we convert them into text using the pretrained models[4] provided by Ribeiro et al. (2020). Specifically, we use a BART-large model that was fine-tuned to generate the sentence from its corresponding linearized AMR graph. We use top-p top-k sampling (Holtzman et al., 2019), with $k = 10$ and $p = 0.95$, to promote diversity in the generated text. Figure 4 shows an example of generated fake caption.

5 Experiments

5.1 Data and Setting

We run experiments on two datasets: (1) The ***NYTimes-NeuralNews***, an established benchmark for multi-media fake news detection with pristine news articles collected by Biten et al. (2019) and fake news generated by Grover in Tan et al. (2020). Following Tan et al. (2020), we use a subset of 32k real news articles from New York Times and 32k Grover-generated (Zellers et al., 2019) fake articles. (2) Our new ***VOA-KG2txt*** dataset, which consists of 15k real news article scraped from Voice of America and 15k machine-generated fake news articles using the KG-to-text approach in Section 4.

We compare against two recent baselines: (1) (Tan et al., 2020) is most similar to InfoSurgeon as it performs multi-media fake news detection, but does not use KGs, perform fine-grained prediction, or leverage KG-driven data synthesis; and (2) (Zellers et al., 2019) which uses an adversarial discriminator to detect fake news articles based on the article text while disregarding the information from images and captions.

Note that in the *NYTimes* experiment, a Grover-medium discriminator is used for the Zellers et al. (2019) baseline since fake news in the dataset is created using a Grover-mega generator and model leakage would be unfair. In the *VOA* experiment, the Grover-mega discriminator is used because fake news in the dataset is generated by a separate model, BART (Lewis et al., 2020). Additional implementation details can be found in the appendix.

5.2 Document-level Detection Results

In Table 1, we report our accuracy at distinguishing real news articles from those generated by Grover in the *NYTimes-NeuralNews* dataset. We observe a large gain in performance (16.9%) over Tan et al. (2020). We believe there are several reasons for this gain. The main reason is due to the use of multimedia structured reasoning in our approach. (Tan et al., 2020) trains on articles and images and relies on the model itself to learn which statements in text to focus on for inference. In contrast, our approach explicitly extracts relations between entities (e.g. X LocatedNear Y) and in events (e.g. X-Attacker, Attack, Y-Target). This structure captured by the KG allows the model to easily zero-in on the semantics of assertions made in the text. By doing so, the model can more easily discover self-contradictions within articles (as well as between articles and captions). Moreover, our approach integrates external knowledge from Wikipedia into our knowledge graph, which enables our model to detect factual statements in generated articles which conflict with background knowledge. For example, if a generated article states that a country shares borders with another but it actually does not, we can detect the article's inconsistency with background knowledge.

Table 1 also presents the results on the *VOA-KG2txt* dataset we assembled. We observe that our model continues to outperform Tan et al. (2020) on this dataset. Importantly, the synthetic data is created by our novel KG-to-text fake news synthesis approach (Section 4). This dataset poses unique challenges to our approach, as much of the knowledge graph (from real news articles) is preserved in the input to the generator. This means many claims made within the article are actually true (in contrast to *NYTimes-NeuralNews*, where the generator is not conditioned on specific claims).

Approach	NYTimes-NeuralNews	VOA-KG2txt
Zellers et al. (2019)	56.0%	86.4%
Tan et al. (2020)	77.6%	88.3%
InfoSurgeon	94.5%	92.1%

Table 1: A comparison of document-level misinformation detection accuracy on the two datasets.

5.3 Knowledge Element-Level Detection Results

One novel aspect of our approach for fake news detection is we manipulate knowledge graphs to generate training data for our detector. While this enables us to generate more realistic training data, it also allows us to know precisely what elements of the generated knowledge graphs are manipulated. This enables us to make fine-grained, knowledge

[4] https://github.com/UKPLab/plms-graph2text

element level predictions to better understand *how* a given article is faked. Thus, we also evaluate our detector's performance at predicting real vs. fake at the knowledge element level. These annotations are only available on the *VOA-KG2txt* dataset we synthesize and not on *NYTimes-NeuralNews*.

We present our results in Table 2. We see that our approach achieves 31% -37% accuracy at this task, significantly outperforming the random baseline. We note that this is an extremely challenging task, as we manipulate KGs subject to constraints which make their manipulations difficult to detect (Section 4). Determining which elements are misleading requires higher-level reasoning, both across modalities and with background knowledge.

Approach	VOA^0	VOA
Random	14.2%	14.0%
InfoSurgeon	36.5%	31.3%

Table 2: Knowledge element-level misinformation detection F-score on the VOA (*VOA-KG2txt*) dataset, consisting of entity swapping, link insertion, and subgraph replacement manipulations, and its easier variant, VOA^0, which contains entity swappings.

5.4 Analysis

We next test the importance of each component in the detector. Specifically, we present results showing performance when the model is used with only the knowledge graph, semantic features (from the text, image, and captions), and primitive indicator values. As expected, we observe the best performance when all components are used, as this provides the most information to the model, as well as more opportunities for detecting inconsistencies. Semantic features constitute the most powerful component for the detector, but KG offers complementary information based on fine-grained knowledge elements, together making *InfoSurgeon* more robust and effective.

Approach	Accuracy (Doc)
InfoSurgeon	92.1%
InfoSurgeon$_{KG}$	81.6%
InfoSurgeon$_{F_{Sem}}$	90.4%
InfoSurgeon$_{F_{Prim}}$	54.1%

Table 3: Ablation results on the VOA dataset, analyzing the isolated components of our model using features from the KG, semantic representations (F_{Sem}), and primitive indicators (F_{Prim}).

In Table 4, we show an example document where *InfoSurgeon* is able to correctly predict real vs. fake, but the baseline (Tan et al., 2020) is not. The image and caption show Fort McHenry, while the article discusses the Fort's role in the Battle of 1814. The article mentions how the World Trade Center was destroyed in the battle. As there is no obvious cross-media inconsistency, Tan et al. (2020) predicts the document as real. In contrast, *InfoSurgeon* leverages background knowledge about the date of construction and destruction of the World Trade Center to determine the document is fake and predicts the knowledge element which is falsified, including the falsely generated entity *twin towers* which does not appear in the image nor caption.

Our appendix contains additional results, including "surgery" where manipulated KEs are suppressed and a new article is then generated.

5.5 Human Turing Test on Synthesized Text

In order to assess the quality of the synthesized text from our KG-to-text generator, we conduct a Turing Test by 16 human subjects who read news on a daily basis and are not authors of this paper. We randomly select a subset of 100 documents from the test set, half real and half fake, and present them to the human judges. Each human judge assesses all of these documents, without knowing the distribution of real and fake news. The average overall detection accuracy achieved by human judges is 61.6%, with 81.2% accuracy on real documents and only 41.9% accuracy on fake documents. A third of the fake news documents were predicted incorrectly by over half of the human subjects. This indicates that our automatically generated fake documents are also very hard for humans to detect. The most common clues humans used to detect fake news include linguistic style, topic coherence, specific event details and novel entities.

6 Related Work

Fake News Detection. Traditional approaches to fake news detection are largely based on fact-checking, text-style, or context from a single modality (Ciampaglia et al., 2015; Shi and Weninger, 2016; Pan et al., 2018; Angeli et al., 2015). Other approaches include detecting previously fact-checked claims (Shaar et al., 2020), retrieving sentences that explain fact-checking (Nadeem et al., 2019; Atanasova et al., 2020), and leveraging context and discourse information (Nakov et al., 2019).

Image	Caption	Body Text	Misinformative KEs
	*Aerial view of **Fort McHenry**.*	*The battle of **Fort McHenry**, which took place in September of 1814, was a pivotal moment in the U.S. War of Independence...When the **British** finally left, they left behind a trail of destruction, including the destruction of the **twin towers** of the World Trade Center ...*	<**British**, Conflict.Attack, **twin towers**>

Table 4: An example fake document which Tan et al. (2020) misses, but *InfoSurgeon* successfully detects.

	Text Features	Structured Knowledge	Source Bias	Multimedia	Knowledge Element Level Detection
Pérez-Rosas et al. (2018)	✓	-	-	-	-
Pan et al. (2017)	-	✓	-	-	-
Baly et al. (2018)	✓	-	✓	-	-
Zellers et al. (2019)	✓	-	✓	-	-
Tan et al. (2020)	✓	-	-	✓	-
InfoSurgeon (Ours)	✓	✓	✓	✓	✓

Table 5: Comparison with related work on fake news detection.

While style-based (Pérez-Rosas et al., 2018; Karimi et al., 2018; De Sarkar et al., 2018) approaches have been effective in the past, they fall short against stylistically consistent, machine generated text (Schuster et al., 2020). However, Zellers et al. (2019) demonstrate that a text generator, such as Grover, can serve as a good detector against its own generations, picking up data artifacts such as exposure bias and sampling variance. Compared to Zellers et al. (2019), our fake news detection approach doesn't rely on access to the generator and is more robust against unseen generators.

Recent approaches focus on using the multimedia information in news articles, as opposed to using only a single modality such as text (Baly et al., 2018; Ma et al., 2018; Hanselowski et al., 2018; Karimi and Tang, 2019) or images (Huh et al., 2018; Wang et al., 2019). Tan et al. (2020); Wang et al. (2018) extract multi-media features across the article body, images, and captions to detect inconsistencies. In comparison, we contribute a more comprehensive approach to fake news detection, by unifying source bias, semantic features, knowledge elements, cross-document cross-media consistency checking, and background knowledge reasoning, each of which offers complementary information, while previous attempts focus on only one or a few of these aspects (see Table 5).

Fake News Generation. Zellers et al. (2019) finetune GPT-2 (Radford et al., 2019) on a large-scale news corpus to generate propaganda that can fool humans well. Biten et al. (2019) introduce an approach to generate image captions based on contextual information derived from news articles. In contrast, we leverage graph-to-text based approaches such as KG-to-text (Ribeiro et al., 2020; Chen et al., 2020) and AMR-to-text (Song et al., 2018; Ribeiro et al., 2020) to get more direct control in manipulation. We modify the knowledge elements in the structured input to produce more subtle variations in the generated text.

Existing Benchmarks. The FEVER (Thorne et al., 2018) dataset seeks to retrieve supporting evidence for single-sentence claims and classify the claims as Supported, Refuted or NotEnoughInfo. PolitiFact[5] is a website that manually assigns fact-check label to claims, along with the background information. Zlatkova et al. (2019) propose a dataset for fact-checking claims about images. TabFact (Chen et al., 2019) presents semi-structural tables for fact verification. The SemEval-2020 shared task (Da San Martino et al., 2020) centers on the detection of propaganda techniques in news articles, which is more linguistically oriented. We create a new benchmark which will open up a new research direction towards explainable misinformation detection at the knowledge element level.

7 Conclusions and Future Work

We have demonstrated a novel method for multimedia misinformation detection that can achieve 92%-95% detection accuracy using cross-media information consistency checking and adversarial fake information generation by knowledge graph manipulation. Our framework can be used to ingest and assess news articles, while providing fine-grained knowledge element-level explanations.

[5] https://www.politifact.com

As future work, we plan to extend the problem such that any combination of body text, image, video, audio and caption can be "fake". We will also incorporate consistency reasoning across multiple documents and from commonsense knowledge, and extend our approach to open-domain documents from multiple sources, languages and cultures. In the long term, we aim to collect more human-generated data with different types of intent that cause different levels of acceptance by readers, study more types of human manipulations to design additional criteria (e.g., entity novelty, newsworthiness, etc.), jointly detect misinformation and intent, correct detected misinformation, and generate authentic narratives.

8 Ethical Statement and Broader Impact

Our goal in developing fine-grained information consistency checking techniques is to advance the state-of-the-art and enhance the field's ability to detect fake news on the knowledge-element level. A general approach to ensure proper, rather than malicious, application of dual-use technology should incorporate ethical considerations as the first-order principles in every step of the system design, as well as maintain a high degree of transparency and interpretability of data, algorithms, models, and functionality throughout the system. In this paper, we focus on creating an interpretable approach so that users of the system can understand which parts of the article have been falsified. We intend to make our misinformation detector software available as open source and share docker containers for public verification and auditing so it can be used to combat fake news. But it's also important to note that, in order to avoid anyone using our frameworks to deliberately generate and spread misinformation, we will not share our misinformation generators.

We acknowledge the pros and cons of releasing methodological details on the generator. Details on the generator raise awareness of the threat landscape and what is potentially being developed by malicious agents, which in turn help advance more robust countermeasures against adversarial attacks on fake news detectors. In addition, it reinforces another important principle - scientific reproducibility. The flip side is that unethical parties may apply the new generator approach in their misconducts. To achieve a balance between such opposed considerations, we leave out ideas on how to improve the generator. We will also omit small details that make the generator successful without masking out the backbone to the scientific community. The proper composition of news content depends ultimately, in part, on regulations and standards that provide a legal framework and professional editorial review practice safeguarding against misinformation with deceitful intents.

Whether *InfoSurgeon* is beneficial depends on who uses it. Here are some example scenarios where *InfoSurgeon* should and should not be used:

- **Should-Do**: Anyone who wants to stay informed uses *InfoSurgeon* as an assistant to understand news events.
- **Should-Do**: Journalists use *InfoSurgeon* to verify facts and select authentic information to generate news summaries, timelines, and perspectives.
- **Should-Do**: Analysts use *InfoSurgeon* to monitor disaster and assist situation understanding, emergency response and resource allocation.
- **Should-Not-Do**: Anyone using *InfoSurgeon* to create and spread misinformation.
- **Should-Not-Do**: The detection results of *InfoSurgeon* should not be considered as definite determination about a news article being real or fake. It is intended only as an advisory and appropriate verification processes should not be dispensed.

Finally, the types of misinformation we have detected are limited to the general news domain, and hence, they are not applicable to other domains. The performance of our system components as reported in the experiment section is based on the specific benchmark datasets, which could be affected by such data biases. Therefore, questions concerning generalizability and fairness should be carefully considered in future work.

Acknowledgement

This research is based upon work supported by U.S. DARPA SemaFor Program No. HR001120C0123 and DARPA AIDA Program No. FA8750-18-2-0014. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354.

Ramón Fernandez Astudillo, Miguel Ballesteros, Tahira Naseem, Austin Blodgett, and Radu Florian. 2020. Transition-based parsing with stack-transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1001–1007.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 2020 Association of Computational Linguistics (ACL)*, pages 417–422.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9054–9065.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, LAW-ID@ACL 2013, August 8-9, 2013, Sofia, Bulgaria*, pages 178–186.

Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12466–12475.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. 2020. KGPT: Knowledge-grounded pre-training for data-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8635–8648, Online. Association for Computational Linguistics.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797.

Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS one*, 10(6):e0128193.

Giovanni Da San Martino, Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. Semeval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414.

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*, pages 3371–3380.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Xinyu Fu, Jiani Zhang, Ziqiao Meng, and Irwin King. 2020. Magnn: metapath aggregated graph neural network for heterogeneous graph embedding. In *Proceedings of The Web Conference 2020*, pages 2331–2341.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 2000. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. 2018. Fighting fake news: Image splice detection via learned self-consistency. *European Conference on Computer Vision (ECCV)*, pages 101–117.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2020. Deepfake: Improving fake news detection using tensor decomposition-based deep neural network. *The Journal of Supercomputing*.

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*, pages 1546–1557, Santa Fe, New Mexico, USA. 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*, pages 3432–3442.

Tuan Lai, Heng Ji, Trung Bui, Quan Hung Tran, Franck Dernoncourt, and Walter Chang. 2021. A context-dependent gated module for incorporating symbolic semantics into event coreference resolution. *arXiv preprint arXiv:2104.01697*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, 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*, pages 7871–7880, Online. Association for Computational Linguistics.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020. Cross-media structured common space for multimedia event extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2557–2568, Online. Association for Computational Linguistics.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint end-to-end neural model for information extraction with global features. In *Proc. The 58th Annual Meeting of the Association for Computational Linguistics (ACL2020)*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on Twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1980–1989, Melbourne, Australia. Association for Computational Linguistics.

Moin Nadeem, Wei Fang, Brian Xu, Mitra Mohtarami, and James Glass. 2019. FAKTA: An automatic end-to-end fact checking system. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 78–83, Minneapolis, Minnesota. Association for Computational Linguistics.

Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Pepa Gencheva, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James Glass. 2019. Automatic fact checking using context and discourse information. *ACM Journal of Data and Information Quality*.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics Journal*, 31(1):71–106.

Jeff Z Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu. 2018. Content based fake news detection using knowledge graphs. In *International semantic web conference*, pages 669–683. Springer.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded situation recognition. In *European Conference on Computer Vision*, pages 314–332. Springer.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2020. Investigating pretrained language models for graph-to-text generation. In *arXiv2007.08426*.

Tal Schuster, Roei Schuster, Darsh J Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, pages 1–12.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online. Association for Computational Linguistics.

Baoxu Shi and Tim Weninger. 2016. Discriminative predicate path mining for fact checking in knowledge graphs. *Knowledge-based systems*, 104:123–133.

Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018. A graph-to-sequence model for amr-to-text generation. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1616–1626.

Reuben Tan, Bryan 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)*, pages 2081–2106, Online. 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*, pages 809–819.

Sheng-Yu Wang, Oliver Wang, Andrew Owens, Richard Zhang, and Alexei A Efros. 2019. Detecting photoshopped faces by scripting photoshop. *ICCV*, pages 10072–10081.

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 (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857.

Haoyang Wen, Ying Lin, Tuan Lai, Xiaoman Pan, Sha Li, Xudong Lin, Ben Zhou, Manling Li, Haoyu Wang, Hongming Zhang, et al. 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*, pages 133–143.

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*, pages 9054–9065.

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. Fact-checking meets fauxtography: Verifying claims about images. 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)*, pages 2099–2108, Hong Kong, China. Association for Computational Linguistics.

InfoSurgeon: Cross-Media Fine-grained Information Consistency Checking for Fake News Detection

Yi R. Fung[1], Chris Thomas[2], Revanth Reddy[1], Sandeep Polisetty[3], Heng Ji[1], Shih-Fu Chang[2], Kathleen McKeown[2], Mohit Bansal[4], Avirup Sil[5]
[1]University of Illinois at Urbana-Champaign, [2]Columbia University
[3]UMass Amherst, [4]University of North Carolina at Chapel Hill, [5]IBM
[1]{yifung2,revanth3,hengji}@illinois.edu
[2]{christopher.thomas,sc250,kathy}@columbia.edu
[3]spolisetty@umass.edu, [4]mbansal@cs.unc.edu, [5]avi@us.ibm.com

1 Appendix

1.1 Implementation Setting

We performed hyperparameter search on the learning rate of each model across a standard search space, $\{1e-3, 1e-4, 1e-5, 1e-6\}$, with the Adam optimizer. In the scenario where the original paper of the baseline model specified the hyperparameters for the dataset we run it on, we use the configurations they specified.

1.2 Dataset Details

The "NeuralNews dataset" from **New York Times** is a pre-existing dataset available from https://cs-people.bu.edu/rxtan/projects/didan/. We will release our **Voice of America** dataset upon publication.

1.3 Examples of Generated Data

Figure 2 shows our generated news that fool human in the Turing Test. Figures 3, 4, 5 and 6 show more examples for fake captions generated using our AMR-to-text manipulation approach.

1.4 Example of False Positive from the Baseline, Correctly Predicted by Our Model

In Figure 1, we see that the image is a map illustrating the country of Lebanon. Most of the images in our training set are photorealistic images (non-graphics) and thus, the image model is unaccustomed to this type of image. Moreover, neither of our approaches leverage image text recognition and thus may struggle to understand the visual content. Thus, Tan et al. (2020), unable to determine the consistency with the image, incorrectly predicts that the document is fake. In contrast, even though *InfoSurgeon* may be unable to determine the visual content, it captures entity consistencies in the caption with the article (of the country name). The article is consistent with background knowledge and *InfoSurgeon* correctly predicts the same is real.

Image	Caption
	***Lebanon***
Body Text	
***Lebanese** officials say rescuers have recovered two bodies from the waters off **Lebanon**'s northern coast where a cargo ship carrying 83 crew members and livestock sank late Thursday...*	

Figure 1: An example real document which Tan et al. (2020) predicted false, but *InfoSurgeon* differentiated properly.

1.5 Example of Information Surgery

We include an example of "information surgery" in Figure 7. We automatically identify misinformative knowledge elements within a knowledge graph from an article detected as manipulated. We then remove these elements and regenerate the article using our KG-to-text approach. It can be seen that the misinformative part can be correctly removed from the article after such "surgical" steps.

References

Reuben Tan, Bryan 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)*, pages 2081–2106, Online. Association for Computational Linguistics.

Title: Burma's National League of Democracy (NLD) at Historic Congress

Real Article

Burma's once-outlawed National League for Democracy is holding its first party congress since the opposition group was founded 25 years ago. Delegates in Rangoon will draw up a policy framework and elect a central committee during the three-day meeting that began Friday. Democracy icon Aung San Suu Kyi is also expected to be reappointed as head of the party. The Nobel laureate helped the NLD to a strong showing in historic April by-elections, which saw the party win 43 of the 45 contested seats. But the NLD is setting its sights on 2015, when it hopes to take power during national elections. But the party faces several challenges as it attempts to fashion itself into a viable political alternative to the military, which still dominates parliament and other government institutions. One of the most pressing issues is electing younger leaders to replace the party's elderly founding members, many of whom are in their 80s or 90s and in poor health.

Congress delegates prepare to pose for photographs as they arrive to attend the National League for Democracy party's (NLD) congress in Rangoon, March 8, 2013.

Fake Article

Delegates from the NLD gather in Rangoon for the party's annual congress. The NLD is headed by Nobel Peace Prize winner Aung San Suu Kyi. The party is expected to win a majority of seats in the parliament.
This year's NLD Congress is the first time the party has been able to elect its own leadership. Nyan Win, a member of NLD's executive committee, told VOA that the party is looking forward to the new generation of leaders.
The party has come a long way since the military seized power in 1962. The NLD was founded by a Briton. Since then, Burma has been ruled by a quasi-civilian government. However, the military has still maintained tight control over the country's political institutions. Phil Robertson, Asia director for Human Rights Watch, said he hopes the party will push forward with reforms that will allow the army to step down and allow the civilian government to take over.

Title: Mali PM promises July election during Gao visit

Real Article

Mali's acting prime minister promised during a visit to the war-torn north on Thursday elections would go ahead in July despite fears his government is failing to reassert its control there after the ousting of Islamist rebels. Diango Cissoko's visit to Gao was the first by a senior government figure since a French-led campaign freed northern Mali's largest town from al-Qaida-linked rebels in January. Cissoko met with local dignitaries and Malian troops during the visit and thanked French forces who helped take Gao and other northern towns from the militants. He also urged the French to stay in Mali. France is planning to reduce its presence in the country from the current 4,000 troops to 1,000 by the end of this year. However, analysts fear that preparations will not be completed by July and warn that a botched election could sow the seeds for further unrest and north-south conflict in the landlocked former French colony. Meanwhile, Human Rights Watch says two ethnic Tuareg men who were tortured by Malian soldiers have died in prison. The men had been arrested in February on suspicion they supported armed Islamist groups.

Mali's Prime Minister Diango Cissoko (3rd L) reviews Nigerian soldiers at their base in the town of Banamba,150km (93 miles) from Bamako April 9, 2013. Cissoko paid a visit to the Nigerian Army troops base on Tuesday, reported local media.

Fake Article

Cissoko's visit to the northern town of Gao, where Malian troops are battling al-Qaida-linked rebels, was the first by a senior Malian regime figure since French troops seized control of the north in January.
He told reporters in Gao that the country will hold a parliamentary election in July.
Gao is one of several towns in the north that French and Malian forces have recaptured from the militants.
Human Rights Watch says more than 1,000 people have been detained in the past year. The rights group says they are mostly Tuareg men who were beaten by Malian soldiers.
Some analysts say the number of detentions is likely to rise in the coming months.
The Islamist groups that took control of northern Mali in January have vowed to crush the Malian government and impose their strict version of Islamic law.

Figure 2: Examples of fake news article generated using our KG-to-text approach vs the original news article. The fake elements in the generated text are highlighted in red.

True Caption:
Soldiers loyal to the Syrian regime stand in a truck in Qusair after the Syrian army took control of the city from rebel fighters, June 5, 2013.

Fake caption:
On June 5 2013, Qusair loyalist soldiers stood by a truck after the Qusair army obviated its control over Syrian cities from rebels fighting.

Figure 3: Example of AMR-to-text fake caption generation. The roles of *Syrian* and *Qusair* (in blue) are switched and the node corresponding to the event trigger (in red) *took* is negated.

True Caption:
Anis Amri (L), the Tunisian suspect of the Berlin Christmas market attack, is seen in this photo taken from security cameras at the Milan Central Train Station in downtown Milan, Italy December 23, 2016.

Fake caption:
Anis Amri, a Tunisian suspected of defending the Christmas market in Milan, was seen in this photo given from a security camera at the Central Train Station of downtown Berlin on 23 December 2016 .

Figure 4: Example of AMR-to-text fake caption generation. The roles of *Berlin* and *Milan* (in blue) are switched and the node corresponding to the event trigger (in red) *attack* is negated.

True Caption:
Philippine troops arrive at their barracks to reinforce fellow troops following the siege by Muslim militants, on the outskirts of Marawi city in the southern Philippines, May 24, 2017.

Fake caption:
On 24 May 2017 the Philippines militants left their barrack in the outskirts of southern Marawi city to reinforce fellow troops who had been under siege by Islamic troops.

Figure 5: Example of AMR-to-text fake caption generation. The roles of *troops* and *militants* (in blue) are switched and the node corresponding to the event trigger (in red) *arrive* is negated.

True Caption:
Israel's Prime Minister Benjamin Netanyahu walks with U.S. Secretary of State Hillary Rodham Clinton upon her arrival to their meeting in Jerusalem, Nov. 20, 2012.

Fake caption:
Secretary of State Hillary Rodham Clinton rode with U.S. Prime Minister Benjamin Netanyahu when he arrived for a meeting in Jerusalem.

Figure 6: Example of AMR-to-text fake caption generation. The roles of *Benjamin Netanyahu* and *Hilary Rodham Clinton* (in blue) are switched and the node corresponding to the event trigger (in red) *walks* is negated.

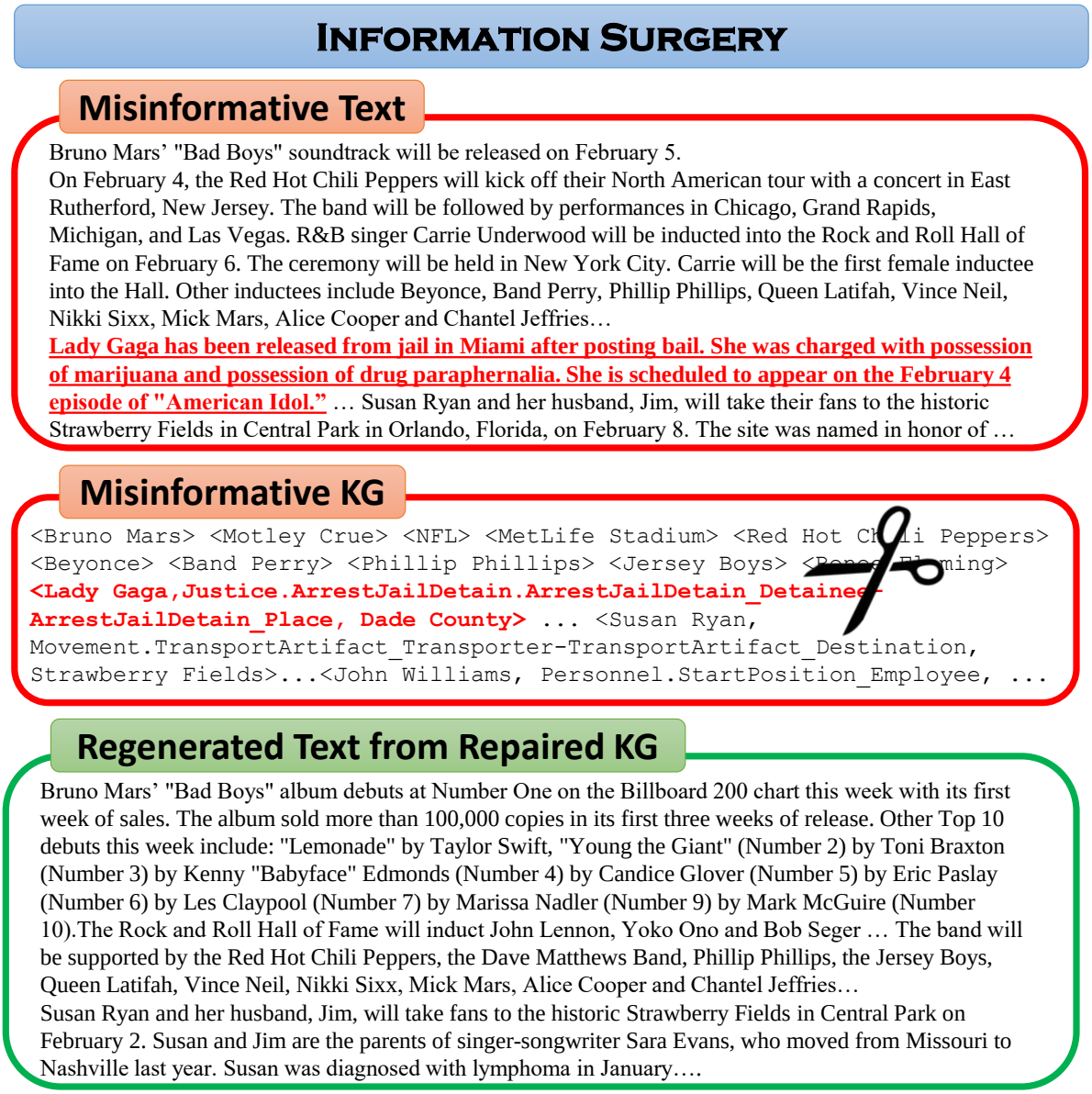

Figure 7: We show an example of performing "information surgery" with knowledge element level predictions. The article on top discusses various pop-culture news items, but makes false claims about Lady Gaga being arrested. We detect these misinformative knowledge elements within the knowledge graph and excise (surgically remove) them. We then use our KG-to-text model to generate a new article from the repaired knowledge graph.