Incorporating External Knowledge for Evidence-based Fact Verification

Anab Maulana Barik
School of Computing
National University of Singapore
anabmaulana@u.nus.edu

Wynne Hsu
Institute of Data Science
National University of Singapore
whsu@comp.nus.edu.sg

Mong Li Lee
NUS Centre for Trusted Internet and Community
leeml@comp.nus.edu.sg

ABSTRACT

Existing fact verification methods employ pre-trained language models such as BERT for the contextual representation of evidence sentences. However, such representations do not take into account commonsense knowledge and these methods often conclude that there is not enough information to predict whether a claim is supported or refuted by the evidence sentences. In this work, we propose a framework called CGAT that incorporates external knowledge from ConceptNet to enrich the contextual representations of evidence sentences. We employ graph attention models to propagate the information among the evidence sentences before predicting the veracity of the claim. Experiment results on the benchmark FEVER dataset and UKP Snopes Corpus indicate that the proposed approach leads to higher accuracy and FEVER score compared to state-of-the-art claim verification methods.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**.

KEYWORDS

fact verification, knowledge graph, language model

ACM Reference Format:
Anab Maulana Barik, Wynne Hsu, and Mong Li Lee. 2022. Incorporating External Knowledge for Evidence-based Fact Verification. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3487553.3524622

1 INTRODUCTION

Advances in the Internet and social media have contaminated digital information with falsehoods [20]. Research in natural language processing has attempted to develop methods to aid the verification of information by predicting whether a set of evidence sentences supports or refutes a claim. For instance, the works in [15, 17] employ the language inference BERT model to verify a claim given the evidence sentences, while GEAR [25] and KGAT [12] use Graph Attention Network (GAT) to make the prediction. We observe that these works do not utilize commonsense knowledge in the claim verification process, and often conclude that there is not enough information to verify the claim. For example, 33.3% of the claims in the benchmark FEVER dataset [18] are deemed to have not enough information by KGAT, and 28% of them supposed to be classified as SUPPORTS or REFUTES.

On closer examination, we discover that for many of these claims, the sentences actually support or refute the claims, as illustrated by the following examples.

Example 1. Consider the pair of claim and evidence sentence:
c1: *"Billboard Dad is a horror film"*
s1: *"Billboard Dad (film) is a 1998 American direct-to-video comedy film"*.

Clearly, the sentence shows that Billboard Dad is a comedy and not a horror film thereby refuting the claim. However, existing works will deem that there is not enough information as they do not know that *"horror"* and *"comedy"* are antonyms.

Example 2. Consider another pair of claim and evidence sentence:
c2: *"Camden, New Jersey is a large human settlement"*
s2: *"Camden is a city in Camden County, New Jersey"*

We see that the claim is supported by the sentence because the phrase *human settlement* is semantically related to *city*. However, existing works will conclude that the claim has not enough evidence as they do not handle phrase-level semantics.

To address the above limitations, we propose a framework called CGAT (for Commonsense Graph Attention Network) that incorporates commonsense knowledge into the claim verification process. Specifically, we utilize the ConceptNet knowledge graph [16] which captures relations between the various words and phrases. For instance, the phrase *"human settlement"* has a relation *"relatedTo"* to *"city"*. CGAT has a graph-based encoder module to inject commonsense knowledge into the representations obtained from some pre-trained language model such as BERT [7] and RoBERTa [11]. Further, we leverage on the structure of ConceptNet [16] to construct a phrase-level graph where each node is a phrase in the claim-sentence pair that can be found in the ConceptNet and two nodes are connected if there exist a path between the corresponding phrases in the ConceptNet. This phrase-level graph is used to obtain the representations of the phrases in the claim-sentence pair.

The proposed framework has a reasoning module to propagate information among the evidence sentences. This is achieved by constructing a fully connected graph where each node is initialized with the knowledge-augmented representation of a claim-sentence pair from the encoder module. This graph forms the input to a graph attention network to refine the representations. The final representations obtained are used to classify the label of a claim whether the claim is supported, refuted, or not enough information. In practice, many of the evidence sentences are retrieved from the web and may be not be relevant to the claim, we design an objective function that considers the relevance of the sentences.

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9130-6/22/04.
https://doi.org/10.1145/3487553.3524622

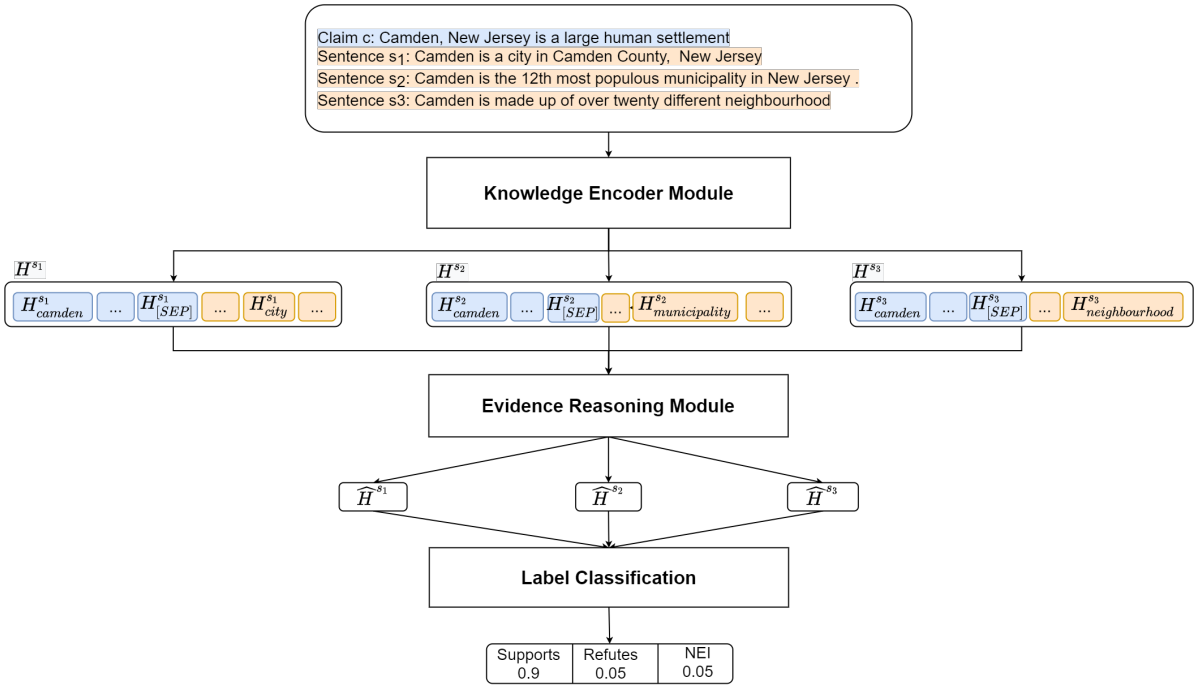

Figure 1: Overview of CGAT.

Experiments on two real world datasets show that the proposed CGAT outperforms state-of-the-art claim verification methods such as GEAR, KGAT and DREAM. Ablation study reveals the incorporating external commonsense knowledge helps to improve the accuracy and facilitate the evidence reasoning process.

2 RELATED WORK

Early works on fact verification typically employ Natural Language Inference (NLI) [1] to classify whether an evidence sentence entails, contradicts, or is neutral with respect to a claim [15, 17]. Subsequent works consider multiple evidence sentences and use graph-based models to propagate the information in these sentences for the verification of claims.

GEAR [25] uses BERT to obtain the representations of claim-sentence pairs which are then modelled as a fully connected graph. Graph attention network (GAT) [19] is employed to propagate the information between nodes in the graph. The final representations for each claim-evidence are obtained from the GAT's last hidden layer to predict the claim label.

KGAT [12] employs BERT to obtain the sentence-level and token-level representations of pairs of claim-evidence sentences and use GAT to propagate information among the neighboring nodes based on the token-level representations. To extract a richer level of interactions between tokens, a kernel-based attention technique [21] is used. The aggregated token-level representations are combined with the sentence-level representations to obtain the final representation of each node for the claim prediction. Substituting the BERT model with stronger language model such as RoBERTa or CorefRoBERTa[23] improves the KGAT's performance further.

DREAM [24] uses semantic role labelling [14] to decompose evidence sentences into spans based on their semantic role (agent or predicate) and construct a semantic graph. XLNet [22] is used to extract the span representations and a graph convolutional network [9] is employed to propagate information among spans to predict whether a claim is supported or refuted.

LOREN [2] and the work in [6] employ a different approach to verify a claim. These works use off-the-shelf Named Entity Recognizer tools to extract central phrases within a claim and verify whether the central phrases are supported by the evidences.

KagNet [10] tries to fill the knowledge gap in question answering by incorporating external knowledge to find paths that can link the concepts mentioned in the question to the concepts in the answers. Similarly, GapQA [8] utilizes ConceptNet to find relationships between the core fact and the answer.

Different from the above works, the proposed CGAT incorporates commonsense into the claim verification process.

3 PROPOSED METHODOLOGY

Given a claim c and a set of retrieved sentences $\{s_1, s_2, \cdots s_N\}$, the goal is to predict the veracity label y of the claim. The label can be SUPPORT, REFUTE, or NEI (for Not Enough Information). Figure 1 gives an overview of the proposed CGAT framework. There are two main components in the proposed framework: the knowledge encoder module and the evidence reasoning module. The encoder module generates a knowledge-aware representation for each claim-sentence pair $<c, s_n>$, $1 \leq n \leq N$. These representations are passed to the reasoning module to predict the claim label y. We discuss the details of the modules in the following subsections.

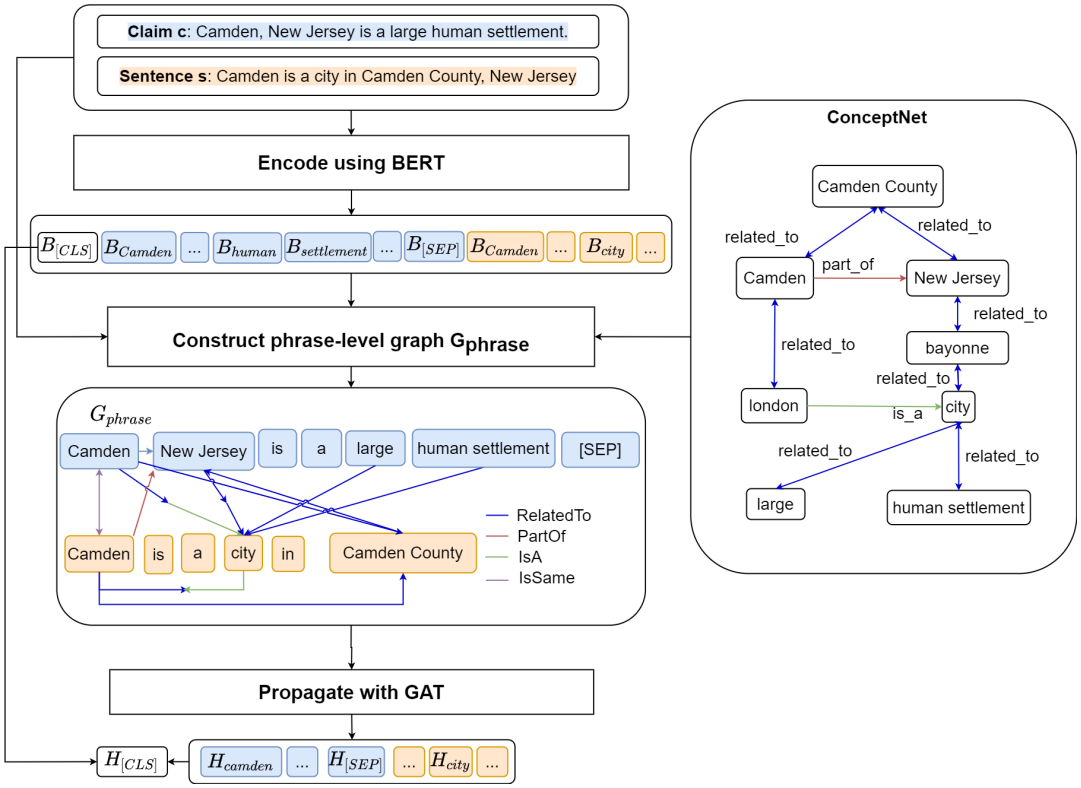

Figure 2: Pipeline of the knowledge encoder module.

3.1 Knowledge Encoder Module

This module creates phrase-level representations so that two phrases that are related semantically will have similar representation. Figure 2 illustrates how the encoder module process a claim-sentence pair. We first concatenate the claim-sentence pair <c, s> from the input as follows:

$$([CLS] + c + [SEP] + Title + [SEP] + s + [SEP])$$

where $Title$ is the title of the document where the sentence s is obtained, $[CLS]$ is the special token marking the start of the claim-sentence pair, and $[SEP]$ are separators.

The output is passed to BERT to obtain the contextual representation B. Then we augment this representation with commonsense knowledge from ConceptNet [16]. ConceptNet [16] is a semantic network comprising of a set of triplets $\langle h, r, t \rangle$ indicating that there is a semantic relation r between concepts h and t.

For each claim-sentence pair <c, s>, we construct a graph G_{phrase} = (V, E) where a node in V is a phrase $p \in \{c, s\}$. This phrase is mapped to some concept node in ConceptNet using entity linking techniques such as Matcher [5]. Specifically, the entities in the ConceptNet form the basis for the Matcher to estabish the mappings between the phrases in the claim or sentence and the entities in the ConceptNet. If a phrase appears multiple times in c or s, then we create a node for each occurrence of the phrase. For a pair of nodes v_i and v_j in G_{phrase}, an edge is created if there exists a k-hop path in the ConceptNet linking the corresponding phrases in v_i and v_j.

For example, we have two nodes for the phrase "Camden", namely $Camden^c$ and $Camden^s$ indicating that the phrase originates from the claim c and evidence s respectively (see Figure 2). The edge between the nodes $Camden^c$ and $New\ Jersey^c$ depicts the semantic relation *relatedTo*. Note that two nodes with the same phrase have an edge depicting the relation *IsSame*.

In the event that there are multiple paths, we will select the shortest path [3]. To break ties among the paths with the shortest length, we will choose the path with relations that occur least frequently. For example, suppose there are 2 paths connecting *tire* and *car* in ConceptNet:

$$(tire \rightarrow RelatedTo \rightarrow car)$$
$$(tire \rightarrow PartOf \rightarrow car)$$

We will choose the second path because the relation *PartOf* only occurs in 0.3% of the relations in ConceptNet compared to the relation *RelatedTo* which occurs 66%. In this way, we aim to capture the more representative relationship between two concepts.

We initialize the nodes in G_{phrase} using the BERT representation. If the phrase in a node consists of multiple words e.g. *"human settlement"*, then we do an average pooling on the representations of the individual words, that is, *"human"* and *"settlement"*. For each edge in G_{phrase}, we initialize it with the BERT representation of the corresponding relation in the ConceptNet. If an edge depict a k-hop path in the ConceptNet, then we concatenate the BERT representations of the k relations in the path.

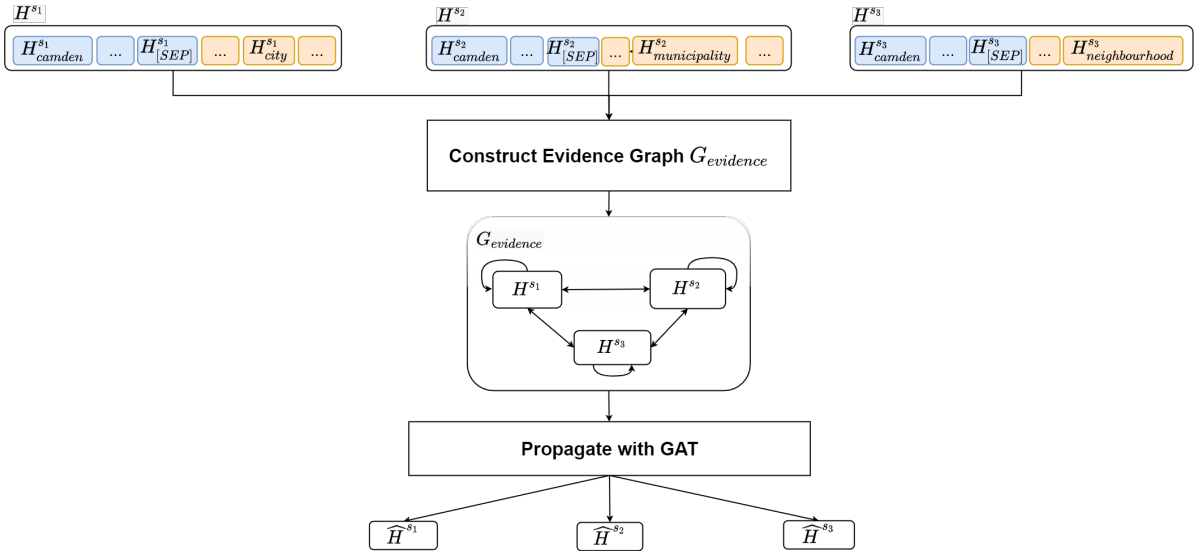

Figure 3: Pipeline of the evidence reasoning module.

After constructing the G_{phrase} graph, we employ Graph Attention Network (GAT) [19] to learn the node representation of v_i taking into account the edge relations from the neighborhood of v_i. Let $\mathcal{N}_i$ be the set of nodes in the neighbourhood of v_i, and h_i^l be the the node representation of v_i at layer l in the GAT. Then the node representation of v_i at layer $l+1$ is given by:

$$h_i^{l+1} = \sigma(\sum_{v_j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}_1 h_j^l + \mathbf{W}_2 e_{ij}) \quad (1)$$

where σ is an activation function, α_{ij} is the importance of v_j to v_i, $\mathbf{W}_1$ and $\mathbf{W}_2$ are $d \times d$ weight matrices for the linear transformation, d is the dimension of the BERT representation, and e_{ij} is the BERT representation of the edge from v_i to v_j.

The final representation of the claim-sentence $<c, s>$ is the concatenation of the node representations in the last layer, denoted as $H^s \in R^{1\times d}$. We update the BERT CLS representation by concatenating the [CLS] token with the average pooling of all the node representations in G_{phrase} before passing it to a linear layer, denoted as $H^s_{[CLS]} \in R^{1\times d}$.

3.2 Evidence Reasoning Module

This module takes as input the final representations of all the claim-sentence pairs to construct an evidence graph $G_{evidence}$ where each node is a claim-sentence pair and the nodes are connected to each other making it a fully connected graph (see Figure 3). Each node i in $G_{evidence}$ is initialized with the final representation H^{s_i} of its claim-sentence pair $<c, s_i>$. Once again, we employ Graph Attention Network (GAT) on the evidence graph to propagate information among the nodes in $G_{evidence}$.

Suppose we have two nodes i and j depicting the claim-sentence pairs $<c, s_i>$ and $<c, s_j>$ with their final representations H^{s_i} and H^{s_j} obtained from the phrase graphs G^i_{phrase} and G^j_{phrase}. We compute the phrase-level attention weight $\boldsymbol{w}^{i\to j}$ based on the cosine similarity between the pairs of nodes in G^i_{phrase} and G^j_{phrase} where the p^{th} entry in $\boldsymbol{w}^{i\to j}$ is given by:

$$\boldsymbol{w}^{i\to j}[p] = \sum_{q \in G^j_{phrase}} \text{sim}(H^{s_i}_p, H^{s_j}_q) \quad (2)$$

We normalize $\boldsymbol{w}^{i\to j}$ using the softmax function before applying it to weight the representation H^{s_i}. Then the information that is propagated from node i to j is given by:

$$\boldsymbol{z}^{i\to j} = H^{s_j}_{[CLS]} \circ (\boldsymbol{w}^{i\to j} \cdot H^{s_i}) \quad (3)$$

where $\circ$ denotes concatenation.

The information propagated is weighted by the sentence-level attention $\beta^{i\to j}$ which is computed as follows:

$$\beta^{i\to j} = \mathbf{W}_3 \cdot (\boldsymbol{z}^{i\to j})^T \quad (4)$$

where $\mathbf{W}_3 \in R^{1\times 2d}$ denotes the weight matrix of a linear transformation and $(\boldsymbol{z}^{i\to j})^T$ is the transpose of $\boldsymbol{z}^{i\to j}$.

We use the softmax function to map the $\beta^{i\to j}$ values to the range $[0, 1]$. With this, we update the representation of H^{s_j} to:

$$\hat{H}^{s_j} = \sum_{i \in G_{evidence}} \beta^{i\to j} \cdot \boldsymbol{z}^{i\to j} \quad (5)$$

3.3 Claim Veracity Classification

Finally, we aggregate all the updated representations using the element-wise max operation as is done in [25]:

$$\boldsymbol{o} = \text{element-max}(\hat{H}^{s_1}, ..., \hat{H}^{s_N}) \quad (6)$$

The result $\boldsymbol{o}$ is passed through a linear layer with three output nodes (one for each label) followed by softmax to obtain the probabilities of SUPPORT, REFUTE or NEI (Not Enough Information). The output with the highest probability is assigned as the veracity label y of the claim.

We train our CGAT by minimising the objective function:

$$L = L_{class} + L_{relevance} \tag{7}$$

The first component L_{class} uses the multi-class cross entropy for the classification accuracy:

$$L_{class} = \text{cross\_entropy}(y^*, y) \tag{8}$$

where y^* is the ground truth veracity of the claim and y is the predicted veracity label.

The second component $L_{relevance}$ uses binary cross entropy for the relevance of each sentence to the claim. We use cosine similarity to compute a relevance score for each sentence s_i to the claim c:

$$r_i = \sum \sum \text{sim}(phrase_c, phrase_{s_i}) \tag{9}$$

where $phrase_c$ and $phrase_{s_i}$ are phrases in the claim c and sentence s_i respectively.

Let $\boldsymbol{r}$ denotes the vector of relevance scores where the i^{th} entry is r_i. We apply the sigmoid function to $\boldsymbol{r}$ so that each $r_i \in [0, 1]$, $1 \leq i \leq N$. Then we have

$$L_{relevance} = binary\_cross\_entropy(\boldsymbol{r}^*, \boldsymbol{r}) \tag{10}$$

where $\boldsymbol{r}^*$ is binary vector where an entry is 1 if the corresponding sentence is the given evidence for supporting or refuting the claim in the dataset, and 0 otherwise.

4 EXPERIMENTS

In this section, we report the results of experiments to evaluate the effectiveness of the proposed approach. The proposed CGAT framework is implemented using PyTorch [13]. We use the following datasets in our experiments:

(1) FEVER [18]. This is the benchmark fact verification dataset consisting of 185,445 claims created based on Wikipedia articles. The claims have been annotated as SUPPORTS, REFUTES, or NOT ENOUGH INFO (NEI). This dataset also indicates the evidences that are relevant to the verification of each claim.
(2) UKP Snopes Corpus [4] This dataset is obtained from the Snopes fact checking website consisting of 5824 claims[1]. The claims, created from heterogeneous web-sources, have been classified as either SUPPORTS, REFUTES, or NOT ENOUGH INFO. However, some of the claims which are labeled as SUPPORTS or REFUTES do not have the associated evidence sentences. As such, we filter out such claims and obtain a cleaned dataset of 3920 claims.

We employ BERT$_{base}$ [7] and RoBERTa$_{large}$ [11] as the pre-trained language model with the maximum length of the sequence set to 130 as in done in [12]. We construct G_{phrase} using 2-hop paths in the ConceptNet as this is sufficient to propagate information between two entities, in other words, we set $k = 2$. The number of GAT layers in the encoder module is set to 2. The hidden size is set to 768 in BERT$_{base}$ and 1024 RoBERTa$_{large}$, the same as the dimension of the pre-trained language model. We train the model using Adam optimizer with a batch size of 4 and a learning rate of 2e-5 with 2 epochs on the FEVER dataset and 50 epochs on UKP Snopes Corpus.

[1] Although the website states 6422 claims, the downloaded file has only 5824 claims

We use the two standard evaluation metrics for the FEVER dataset, namely label accuracy and FEVER score. Label accuracy measures the correctness of the claim veracity labels without considering the relevance of the retrieved evidences. FEVER score measures the correctness of the claim veracity labels using only the set of relevant evidences. For the UKP Snopes corpus, there is no FEVER score since the dataset does not annotate whether the retrieved evidences are relevant. Besides label accuracy, we also measure the recall and F1 score. We record the mean and standard deviation of the results over five runs.

4.1 Comparative Study

We compare our CGAT with the following methods:

- GEAR [25]. This method employs BERT to obtain the representation of each claim-sentence pair, and then utilizes GAT attention mechanism to aggregate the evidence before giving the prediction.
- KGAT [12]. This version of KGAT employs RoBERTa$_{large}$ to obtain the claim-sentence pair representation before utilizing a fine-grained Kernel GAT to aggregate the evidence for the claim prediction.
- DREAM [24]. This method uses semantic role labeling to chunk the sentences into words/phrases and construct a semantic graph. Then it uses XLNet to obtain the contextual representation of the words/phrases, and propagate the information using a graph convolution network. The information is aggregated by a GAT before the final prediction.

Table 1 shows the label accuracy and FEVER score for the various methods on the FEVER test set. We see that although CGAT has comparable label accuracy as DREAM, it achieves the highest FEVER score, demonstrating that augmenting the evidence sentences with commonsense knowledge from the ConceptNet increases the inference ability of CGAT. Table 2 shows the results on the original and cleaned UKP Snopes datasets. We observe that CGAT outperforms the other methods by a large margin for all the metrics, indicating its robustness. Among the claims that KGAT miss-classified as NOT ENOUGH INFORMATION, CGAT_RoBERTa is able to utilize ConceptNet to fill the knowledge gap between the claim and evidence sentences and correctly re-classify 56% claims in the cleaned UKP Snopes dataset and 25% of the claims in FEVER test set to SUPPORT or REFUTE.

4.2 Ablation Study

We use the lightweight CGAT_BERT model on the FEVER development set and the UKP Snopes cleaned set for our ablation study. We implement the following variants of CGAT_BERT:

(1) CGAT_BERT without Evidence Reasoning. This variant only uses the Knowledge Encoder Module and skips the Evidence Reasoning Module. To obtain the probability of each label, the model performs element-wise max pooling on all [CLS] representations from the Knowledge Encoder Module, and the result is passed through a linear layer with three output nodes followed by a softmax.
(2) CGAT_BERT without ConceptNet. This variant skips the construction of G_{phrase} graphs, and uses the pre-trained BERT language model to initialize the nodes in $G_{evidence}$.

Table 1: Results on FEVER dataset.

Model	Development Set		Test Set	
	Label Accuracy	FEVER Score	Label Accuracy	FEVER Score
GEAR	74.84	70.69	71.60	67.10
KGAT	78.69	76.11	74.07	70.38
DREAM	-	-	**76.85**	70.60
CGAT_BERT	78.08	76.08	73.29	70.05
CGAT_RoBERTa	**80.64**	**78.46**	76.39	**73.15**

Table 2: Results on UKP Snopes Dataset.

Model	Original			Cleaned		
	Label Accuracy	Recall	F1	Label Accuracy	Recall	F1
GEAR	62.43±0	59.29±0	51.35±0	75.89±0	66.6±0	64.52±0
KGAT	68.77±0.94	65.86±1.79	62.05±1.69	81.84±1.28	73.43±0.93	72.65±2.49
DREAM	67.84±1.77	66.21±1.09	63.89±0.87	82.04±0.62	77.15±1.14	77.67±1.04
CGAT_BERT	70.82±0.74	71.92±0.46	70.08±0.65	84.99±1.33	81.11±1.03	81.74±0.97
CGAT_RoBERTa	**73.94±0.46**	**77.21±0.80**	**74.77±0.47**	**86.70±1.47**	**86.88±0.79**	**85.92±0.65**

Table 3: Results of Ablation Study.

Model	FEVER Development Set		UKP Snopes Cleaned Set		
	Label Accuracy	FEVER Score	Label Accuracy	Recall	F1
CGAT_BERT without Evidence Reasoning	77.21±0.64	75.18±0.65	82.06±1.43	79.00±1.47	79.05±0.33
CGAT_BERT without ConceptNet	77.61±0.40	75.59±0.38	82.37±1.31	79.06±1.13	79.56±0.14
CGAT_BERT without Relevance Loss	78.00±0.06	75.81±0.11	84.34±1.21	80.66±0.41	81.23±0.81
CGAT_BERT	**78.05±0.05**	**76.03±0.07**	**84.99±1.33**	**81.11±1.03**	**81.74±0.97**

(3) CGAT_BERT without Relevance Loss. Here, the CGAT model is trained using the cross entropy loss in the training objective function.

Table 3 shows the ablation results. For the FEVER Development set, we observe the drop in label accuracy and FEVER score is the greatest when CGAT does not incorporate the evidence reasoning module indicating the importance of this module in the verification process. This is followed by CGAT without external knowledge from ConceptNet. The decrease in performance is the smallest when we do not include the relevance loss in the objective function. Similar trends is observed on the UKP Snopes Cleaned dataset.

4.3 Case Studies

In this section, we highlight some sample claims in FEVER and UKP Snopes to show that incorporating knowledge from ConceptNet enables CGAT make correct predictions by filling in the knowledge gap between a claim and the evidence sentences.

4.3.1 Correctly predicted claims in FEVER. Claims C1 and C2 in Table 4 are predicted as having not enough information (NEI) by KGAT, while the proposed CGAT is able to provide the correct predictions. For claim C1, the ConceptNet has a path

(earthling → FormOf → Antonym → alien)

indicating an antonym relationship between "earthling" in the claim and "alien" in sentence S1. This has enriched their corresponding BERT representations as the G_{phrase} graph constructed provides the contextual information taking into account the edge relationship antonym. Together with the relatively high relevance of S1 to the claim, CGAT is able to conclude that the claim C1 is refuted.

For claim C2, the ConceptNet has the phrases "international organization" and "United Nations". This enables CGAT to create nodes for these phrases instead of having nodes for the individual words. Further, there is a path

(United Nations → IsA → world organization → Synonym → international organization)

indicating the relationships between the two phrases. With this, CGAT is able to infer that S1 and S2 support the claims and correctly label the veracity of C2.

4.3.2 Correctly predicted claims in UKP Snopes. Similarly, the claims C3 and C4 in Table 5 are correctly predicted as being refuted by CGAT because of the following paths in ConceptNet linking the words "two", "wives", "polygamy" and "marriages" as well as the words "killing", "burying", and "execution":

- *(polygamy → IsA → marriage → RelatedTo → two)*
  (polygamy → RelatedTo → marry → wives)
  (polygamous → RelatedTo → polygamy → IsA → marriage)
  (polygamous → RelatedTo → monogamous)
  (wives → IsA → spouse)
  (immigrants → Antonym → citizen)

Table 4: Sample claims in FEVER that are correctly predicted by CGAT. Value in bracket depicts the relevance score.

	Claim	Evidence	Prediction
C1	Pearl (Steven Universe) is a fictional earthling being.	S1: Pearl is a " Gem ", a fictional alien being that exists as a magical gemstone projecting a holographic body. (0.45) S2: Pearl is a fictional character from the 2013 animated series Steven Universe, created by Rebecca Sugar. (0.31) S3: She is portrayed as a loving, gentle and delicate character, who acts as a motherly figure for Steven. (0.06) S4: However, she also tends to be overprotective with him and has low self esteem... (0.04) S5: It is the coming of age story of a young boy named Steven Universe, who lives in the fictional town of Beach City with the " Crystal Gems" Pearl, Garnet, and Amethyst, three magical humanoid aliens. (0.12)	Ground truth: REFUTES KGAT: NEI CGAT without ConceptNet: NEI CGAT: REFUTES
C2	Ukrainian Soviet Socialist Republic was in an international organization.	S1: Ukrainian SSR was a founding member of the United Nations, although it was legally represented by the All Union state in its affairs with countries outside of the Soviet Union. (0.45) S2: The United Nations LRB UN RRB is an inter-governmental organization to promote international co-operation and to create and maintain international order. (0.31) S3: Ukrayina was one of the constituent republics of the Soviet Union from its inception in 1922 to its breakup in 1991. (0.06) S4: The Ukrainian Soviet Socialist Republic (Ukrainian SSR or UkrSSR or UkSSR), commonly referred to in English as Ukraine LRB LSB ... RSB. (0.04) S5: The Ukrainian SSR was situated in Eastern Europe to the north of the Black Sea, bordered by the Soviet republics of Moldavia, Byelorussia, and the Russian SFSR. (0.12)	Ground truth: SUPPORTS KGAT: NEI CGAT without ConceptNet: NEI CGAT: SUPPORTS

- *(killing → HasLastSubEvent → death → execution)*
  (burying → RelatedTo → killing)
  (burying → murder → execution)

4.3.3 Incorrectly predicted claims. Table 6 gives two sample claims where both KGAT and CGAT make the wrong predictions. Claim C5 is from the FEVER dataset, while C6 is from UK Snopes. KGAT focuses on the token-level representations. Although Syria is different from Iran, KGAT still infers that claim C5 is supported due to the large number of overlapping tokens between the claim and evidence sentences. On the other hand, CGAT recognizes that there is a relationship between Syria and Iran given the path in ConceptNet:

(Syria → PartOf → asia → RelatedTo → Iran)

However, in this context, the general knowledge that the two countries Iran and Syria are in the same continent is not helpful and leads to the wrong conclusion.

For Claim C6, we see that both KGAT and CGAT conclude that there is not enough information. At the token-level, there is no overlapping words. Further, the ConceptNet does not contain phrases "impaled with a fork" and "the fork was through the nose" as these phrases are uncommon, leading to the wrong prediction.

5 CONCLUSION

We have proposed a framework called CGAT to incorporate external knowledge to the evidence-based fact verification process. We utilized the relations and structure in the ConceptNet to enrich the phrase-level representations of the claims and evidence sentences. With this, we constructed an evidence graph and employed graph attention networks to propagate the information among the evidence sentences before predicting the veracity of the claim. We have also introduced the relevance loss component in the objective function to handle evidences that have different degrees of match to the claim. Experiment results on benchmark datasets demonstrated the effectiveness of CGAT to increase the label accuracy and FEVER score over state-of-the-art claim verification methods.

Despite the advantages of our proposed approach, there is still room for improvements as indicated in the case studies. Future work includes investigating ways to take into account geographical and temporal information in the verification process.

Table 5: Sample claims in UK Snopes that are correctly predicted by CGAT. Value in bracket depicts the relevance score.

	Claim	Evidence	Predictions
C3	Canadian immigrants with two wives receive a host of government benefits upon their their arrival.	S1: Polygamy is illegal in Canada, and therefore multiple marriages are not recognized under Canada's immigration laws. (0.25) S2: This means that a permanent resident or Canadian citizen can only immigrate with one spouse after having dissolved other marriages to "convert" their polygamous marriage to a monogamous one. (0.24) S3: Immigration, Refugees and Citizenship Canada (IRCC) has advised the United Nations Refugee Agency that individuals in a polygamous marriage should not be referred for resettlement to Canada. (0.24) S4: As well, IRCC officers assess privately sponsored refugee cases against Canada's immigration laws, including monogamous marriage requirements. (0.25)	Ground truth: REFUTES KGAT: NEI CGAT without ConceptNet: NEI CGAT: REFUTES
C4	U.S. General John J. Pershing effectively discouraged Muslim terrorists in the Philippines by killing them and burying their bodies along with those of pigs.	S1: But the story is not true. (0.003) S2: There was no mass execution led by Pershing. (0.86) S3: That is a rumor created on the Internet. (0.13) S4: The Tribune article says Pershing sprinkled some prisoners with pig's blood, which the Juramentados believed would condemn them for eternity. (0.006) S5: But then Pershing let the prisoners go. (0.001)	Ground truth: REFUTES KGAT: REFUTES CGAT without ConceptNet: REFUTES CGAT: REFUTES

Table 6: Sample claims that are wrongly predicted by CGAT. Value in bracket depicts the relevance score.

	Claim	Evidence	Prediction
C5	Hezbollah received financial support from Syria.	S1: Hezbollah receives military training, weapons, and financial support from Iran, and political support from Syria. (0.72) S2: Hezbollah was conceived by Muslim clerics and funded by Iran primarily to harass the Israeli occupation. (0.26) S3: Since 2012, Hezbollah has helped the Syrian government during the Syrian civil war in its fight against the Syrian opposition, which Hezbollah has described as a Zionist plot and a " Wahhabi Zionist conspiracy" to destroy its alliance with Assad against Israel. (0.05) S4: Hezbollah maintains strong support among Lebanon's Shi'a population, while Sunnis have disagreed with the group's agenda. (0.03) S5: After the Israeli invasion of Lebanon in 1982 in support of the Free Lebanon State, Israel occupied a strip of south Lebanon, which was controlled by the South Lebanon Army LRB SLA RRB, a Lebanese Christian militia supported by Israel. (0.01)	Ground truth: REFUTES KGAT: SUPPORTS CGAT without ConceptNet: SUPPORTS CGAT: SUPPORTS
C6	Photographs show a boy whose nose has been impaled with a fork.	S1: When the waiter picked him up from under the table the fork was through his nose. (0.47) S2: The one picture is from the ER and the other picture is two days later at home. (0.53)	Ground truth: SUPPORTS KGAT: NEI CGAT without ConceptNet: NEI CGAT: NEI

REFERENCES

[1] Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 632–642.

[2] Jiangjie Chen, Qiaoben Bao, Jiaze Chen, Changzhi Sun, Hao Zhou, Yanghua Xiao, and Lei Li. 2020. LOREN: Logic Enhanced Neural Reasoning for Fact Verification. *arXiv preprint arXiv:2012.13577* (2020).

[3] Kshitij Fadnis, Kartik Talamadupula, Pavan Kapanipathi, Haque Ishfaq, Salim Roukos, and Achille Fokoue. 2019. Path-Based Contextualization of Knowledge Graphs for Textual Entailment. (2019).

[4] Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A Richly Annotated Corpus for Different Tasks in Automated Fact-Checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. 493–503.

[5] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017). To appear.

[6] Mayank Jobanputra. 2019. Unsupervised Question Answering for Fact-Checking. *EMNLP 2019* (2019), 52.

[7] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.

[8] Tushar Khot, Ashish Sabharwal, and Peter Clark. 2019. What's Missing: A Knowledge Gap Guided Approach for Multi-hop Question Answering. 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)*. Association for Computational Linguistics, Hong Kong, China, 2814–2828. https://doi.org/10.18653/v1/D19-1281

[9] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[10] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. 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)*. Association for Computational Linguistics, Hong Kong, China, 2829–2839. https://doi.org/10.18653/v1/D19-1282

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. (2019).

[12] Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained Fact Verification with Kernel Graph Attention Network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7342–7351.

[13] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: an imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 8026–8037.

[14] Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255* (2019).

[15] A Soleimani, C Monz, and M Worring. 2020. BERT for Evidence Retrieval and Claim Verification. *Advances in Information Retrieval* 12036 (2020), 359–366.

[16] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

[17] Dominik Stammbach and Guenter Neumann. 2019. Team DOMLIN: Exploiting evidence enhancement for the FEVER shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*. 105–109.

[18] 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, Volume 1 (Long Papers)*. 809–819.

[19] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.

[20] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[21] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval*. 55–64.

[22] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).

[23] Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential Reasoning Learning for Language Representation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 7170–7186.

[24] Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning Over Semantic-Level Graph for Fact Checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6170–6180.

[25] Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based Evidence Aggregating and Reasoning for Fact Verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 892–901.