SISER: Semantic-Infused Selective Graph Reasoning for Fact Verification

Eunhwan Park[1♠], **Jong-Hyeon Lee**[2♠*], **Donghyeon Jeon**[3], **Seonhoon Kim**[3], **Inho Kang**[3], **Seung-Hoon Na**[1†]
[1]Jeonbuk National University, [2]NCSOFT, [3]NAVER Corporation
{judepark, nash}@jbnu.ac.kr, leejh1230@ncsoft.com
{donghyeon.jeon, seonhoon.kim, once.ihkang}@navercorp.com

Abstract

This study proposes **S**emantic-**I**nfused **SE**lective Graph **R**easoning (SISER) for fact verification, which newly presents semantic-level graph reasoning and injects its reasoning-enhanced representation into other types of graph-based and sequence-based reasoning methods. SISER combines three reasoning types: 1) *semantic*-level graph reasoning, which uses a semantic graph from evidence sentences, whose nodes are elements of a triple – <Subject, Verb, Object>, 2) "semantic-infused" *sentence*-level "selective" graph reasoning, which combine semantic-level and sentence-level representations and perform graph reasoning in a selective manner using the node selection mechanism, and 3) *sequence* reasoning, which concatenates all evidence sentences and performs attention-based reasoning. Experiment results on a large-scale dataset for Fact Extraction and VERification (FEVER) show that SISER outperforms the previous graph-based approaches and achieves state-of-the-art performance.

1 Introduction

An ever-increasing number of unconfirmed false or misleading information spread on various social media platforms has motivated the verification of textual information, referred to as *fact verification*. FEVER (Thorne et al., 2018a) presented a large dataset for fact verification, initiating a shared task that aims to automatically classify a human-generated claim into *'Supported'*, *'Refuted'*, or *'Not Enough Info'* based on retrieved evidence sentences from Wikipedia[1].

Claim verification, the final step of fact verification, is viewed as a task of natural language inference (NLI) (Angeli and Manning, 2014). Specifically, the NLI task for claim verification is formulated as the *set-to-sentence* entailment of inferring whether a claim (as the hypothesis) is logically "entailed" from a set of retrieved evidence sentences (as the premise).

Recently, *graph reasoning* for claim verification has been extensively explored (Zhou et al., 2019; Liu et al., 2020; Zhong et al., 2020), which creates a graph whose nodes are semantic units extracted from a set of evidence sentences or a claim, and applies graph neural networks (GNNs) such as (Veličković et al., 2018; Kipf and Welling, 2017) to infer the entailment relationship. However, graph reasoning may be somehow restricted to *unit-biased reasoning*, when relying on a single type of semantic unit for nodes of a graph, such as sentences, entities, or words, meaning that the semantic interaction between claim and evidence is restricted to a single graph type and does not go beyond the coverage of the "given" semantic units. In addition, graph reasoning may suffer from *over-smoothing* inherited from GNNs (Gasteiger et al., 2019; Zhao and Akoglu, 2020; Chen et al., 2020a; Rong et al., 2020), likely causing all node representations to converge to a stationary point at the extreme, as reported by (Li et al., 2018).

To address these limitations of graph reasoning, this study proposes **SISER** – **S**emantic-**I**nfused **SE**lective graph **R**easoning) for fact verification by extensively exploiting additional semantic units for graph reasoning and integrating semantic-level reasoning with sequence reasoning and "selective" graph reasoning. SISER combines the following three types of reasoning:

- ***Semantic*-level graph reasoning** applies GNNs to a "semantic graph" whose nodes are elements of <Subject, Verb, Object> that appear in evidence sentences. Provided fine-grained semantic granularity, it is expected that the use of semantic elements would be helpful to effectively induce their own dis-

[†]Corresponding author
[♠]Equal contribution
[*]This work was done during his M.S. degree at JBNU.
[1]https://competitions.codalab.org/competitions/18814

Proceedings of the 29th International Conference on Computational Linguistics, pages 1367–1378
October 12–17, 2022.

tinct representations useful for claim verification, compared to sentence-level representations.

- **Semantic-infused *sentence*-level selective graph reasoning** combines semantic- and sentence-level representations and performs selective graph reasoning equipped with a node selection mechanism. Motivated by variants of GNNs (Gasteiger et al., 2019; Zhao and Akoglu, 2020; Chen et al., 2020a; Rong et al., 2020) to handle oversmoothing issues, we further provide "selective" graph reasoning where a subset of nodes is "selected" using the *node selection mechanism* and only these selected nodes participate in graph reasoning[2]. It is expected that the node selection mechanism can alleviate oversmoothing by breaking full connectivity.
- ***Sequence* reasoning**, concatenates a claim and all evidence sentences and performs self-attention over the concatenated long sequence. As in (Kruengkrai et al., 2021), it is expected that sequence reasoning shows stable performance, without suffering from the inherent problems of GNNs.

Furthermore, we newly apply *prompt-based fine-tuning* (Schick and Schütze, 2021a; Gao et al., 2021) by reformulating the fact verification task as a masked language modeling problem, where a *label word* is generated on a given prompt with a task-specific *template*. To the best of our knowledge, this is the first attempt to use semantic-level 'selective' graph reasoning and prompt-based fine-tuning for the fact verification task.

Our contributions are summarized as follows: 1) We propose SISER, which consists primarily of semantic-level reasoning and semantic-infused selective graph reasoning using the node selection mechanism for fact verification; 2) We present the initial work of adopting prompt-based fine-tuning for claim verification; 3) The proposed SISER shows state-of-the-art performance in the FEVER dataset.

2 Related Work

2.1 Fact Verification Systems

Sequence Reasoning

The baseline system (Thorne et al., 2018a) concatenates all retrieved evidence sentences and then feeds the concatenated evidence and a claim into a pretrained language model as an early sequence reasoning method. The studies of (Hanselowski et al., 2018; Hidey and Diab, 2018) proposed adapting the enhanced sequential inference model (ESIM) (Chen et al., 2017) to measure the semantic relatedness between a claim and evidence. Nie et al. (2019) proposed a carefully designed neural semantic matching network (NSMN), which is a modification of the enhanced sequential inference model. Unlike treating the fact verification task as an NLI task, LOREN (Chen et al., 2020b) proposed decomposing the verification of the entire claim at the phrase level, where the veracity of the phrases serves as explanations and can be aggregated into the final verdict according to logical rules. More recently, MLA (Kruengkrai et al., 2021) argued that graph reasoning may be unnecessary for a claim verification task, proposing *multi-level* sequence reasoning that consists of $\{\mathsf{token}, \mathsf{sentence}\}$-level self-attention (Vaswani et al., 2017).

Graph Reasoning

In contrast to ESIM, NSMN, and LOREN, GEAR (Zhou et al., 2019) proposed graph-based evidence reasoning using GNNs, which conducts reasoning and aggregation over claim-evidence pairs under an evidence graph (Veličković et al., 2018; Kipf and Welling, 2017). Similarly, KGAT (Liu et al., 2020) proposed the use of a semantic-level graph for fine-grained evidence reasoning that uses a kernel-based graph attention mechanism to properly propagate information between nodes. Unlike KGAT, DREAM (Zhong et al., 2020) considered a word span obtained by semantic role labeling (SRL) as a node in the graph and employed XLNet (Yang et al., 2019) as a pretrained language model. In contrast to existing graph reasoning studies that rely on sentence-level or semantic-level graphs, SISER extensively uses "heterogeneous" graphs and fuses different types of reasoning-enhanced representations, going beyond the limitation of using only a single type of reasoning.

2.2 Prompt-based Fine-tuning

PET introduces prompt-based learning, which treats a downstream task as a masked language modeling problem and performs gradient-based fine-tuning (Schick and Schütze, 2021a,b). Employing prompt-based fine-tuning can reduce the gap between pre-training and fine-tuning, which

[2]Here, the selection process is random but parameterized by neural models.

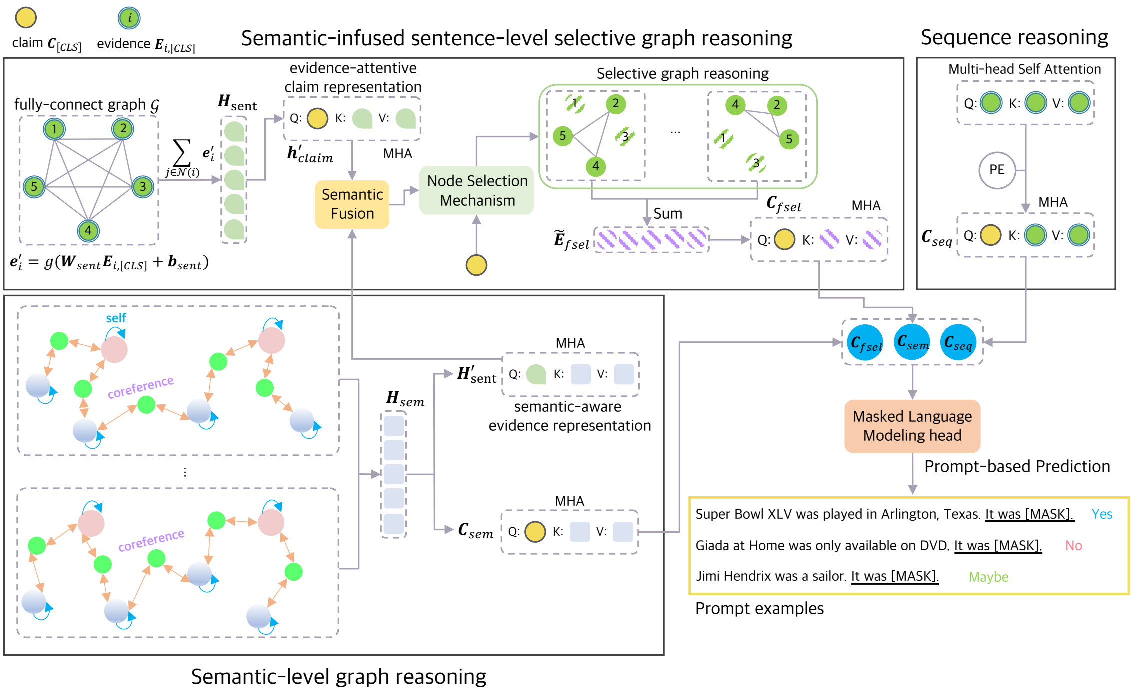

Figure 1: A neural architecture of the proposed SISER: 1) The semantic-level graph reasoning is performed using R-GCN on a semantic graph constructed using the Levi graph transformation to generate the semantic-level node representation $\boldsymbol{H}_{sem}$ (Eq. (2)), which is used to induce the *semantic-aware evidence representation* $\boldsymbol{H}'_{sent}$ (Eq. (5)). 2) The semantic-infused sentence-level selective graph reasoning performs the the selective graph reasoning on a sub-graph resulting from the *node selection mechanism* based on the semantic-fused representation of $\boldsymbol{h}'_{claim}$ (Eq. (5)) and $\boldsymbol{H}'_{sent}$ to generate $\tilde{\boldsymbol{E}}_{fsel}$ (Eq. (10)). 3) The sequence reasoning performs MHA on m evidence representations $\boldsymbol{E}_{seq}$ (Eq. (11)) to obtain $\boldsymbol{H}_{seq}$. 4) The prompt-based claim verification performs the prediction of label-verbalized words at [MASK]'s position on the fused semantic-attentive claim representations $\boldsymbol{H}$ induced from $\boldsymbol{C}_{fsel}, \boldsymbol{C}_{sem}, \boldsymbol{C}_{seq}$ as in Eq. (12).

makes it effective for various tasks. Inspired by PET, LM-BFF (Gao et al., 2021) introduced the adaptation of prompt-based learning to few-shot fine-tuning. Moreover, this study proposed an automatic prompt search method to resolve the difficulty of finding the optimal task-specific template. P^3 Ranker (Hu et al., 2022) proposed a pre-trained, prompt-learned, pre-finetuned neural ranker that employs prompt-based learning to convert the ranking task into pre-training and uses pre-finetuning. (Ding et al., 2021) introduced adapting prompt learning into an entity typing task in several scenarios (e.g., fully supervised, few-shot, zero-shot), which shows the possibility of employing prompt-based learning in fully supervised scenarios. Unlike several methods that employ prompt-based learning in a few-shot scenario, we adapt prompt-based learning in a fully supervised scenario.

3 Proposed Approach

Figure 1 shows the overall neural architecture of the proposed SISER model, which combines three types of reasoning: i.e., semantic-level graph reasoning; semantic-infused sentence-level selective graph reasoning; and sequence reasoning. This section presents details of the three reasoning methods.

3.1 Initial Representation of Claim and Evidences

Suppose that a claim c and a set of retrieved evidence sentences $\{\mathsf{e}_1, \cdots, \mathsf{e}_m\}$ are presented for a fact verification task, where m is the number of evidence sentences and PLM refers to the encoder of a pretrained language model such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). Feeding a claim-evidence pair $(\mathsf{c}, \mathsf{e}_i)$ for the i-th evidence sentence and claim c into PLM, we obtain $\boldsymbol{E}_i$ and $\boldsymbol{C}$ as evidence and claim representa-

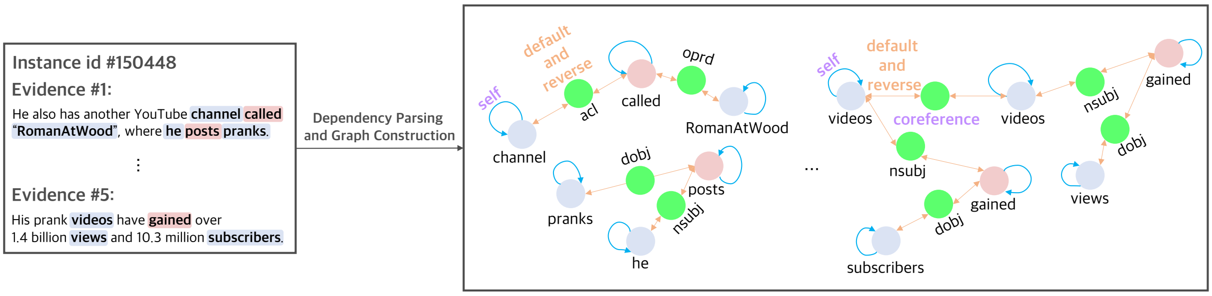

Figure 2: An illustration of constructing a semantic graph for sentence-level graph reasoning, motivated by the procedure of (Beck et al., 2018): 1) a (large) *dependency graph* is first obtained by applying the Spacy's syntactic parser (Honnibal and Montani, 2017) and the NeuralCoref's coreference resolution to m evidence sentences where each occurrence of a word is treated differently with its contextual representation. When two mentions are coreferent, their head words are connected by the "coreference" relation. 2) The dependency graph is then transformed to a *semantic graph* using the Levi graph transformation of (Beck et al., 2018) by including dependency labels as a node set with three types of edge labels – $\{\mathsf{default}, \mathsf{reverse}, \mathsf{self}\}$.

tions as follows:

$$\begin{aligned} \boldsymbol{E}_i &= \mathsf{PLM}\left(\mathsf{c}, \mathsf{e}_i\right) \in \mathbb{R}^{(|\mathsf{c}|+|\mathsf{e}_i|)\times d_{model}}, \\ \boldsymbol{C} &= \mathsf{PLM}\left(\mathsf{c}\right) \in \mathbb{R}^{|\mathsf{c}|\times d_{model}}, \end{aligned} \tag{1}$$

where $|\mathsf{x}|$ is the length of sequence x, and d_{model} is the dimensionality of PLM. Let $\boldsymbol{E}_{i,[\mathsf{CLS}]} \in \mathbb{R}^{d_{model}}$ and $\boldsymbol{C}_{[\mathsf{CLS}]} \in \mathbb{R}^{d_{model}}$ be representations of [CLS] tokens for e_i and c, respectively.

3.2 Semantic-level Graph Reasoning

Our semantic-level reasoning is similar to the work of (Zhong et al., 2020), but differs in using semantic units and types of GNNs, as described below.

3.2.1 Semantic Graph

Similar to (Beck et al., 2018), we construct a semantic graph based on graph transformation, starting from a dependency graph. More specifically, we first obtain a *dependency graph* $\mathcal{G}_{dep} = (\mathcal{V}_{dep}, \mathcal{E}_{dep})$, resulting from m by parsing all m evidence sentences using Spacy's syntactic parser (Honnibal and Montani, 2017)[3] and NeuralCoref's coreference resolution[4], where $\mathcal{V}_{dep}$ is a set of "words" that appear in m evidence sentences and $\mathcal{E}_{dep}$ is a set of dependency-labeled edges. When two mentions are connected by a coreference link, the "coreference" relation is appended between their head words. It should be noted that when a word occurs multiple times in m evidence sentences, we treat each occurrence differently by using their contextual representations (i.e., the span representations) as the elementary semantic representations.

We then convert $\mathcal{G}_{dep}$ into a *semantic graph* $\mathcal{G}_{sem} = (\mathcal{V}_{sem}, \mathcal{E}_{sem})$, a Levi Graph based on the graph transformation of (Beck et al., 2018; Cheng et al., 2020; Huang et al., 2021), where $\mathcal{V}_{sem}$ is a combined set of words and dependency relations that appear in m evidence sentences, and $\mathcal{E}_{sem}$ is a set of type-labeled edges whose labels are taken from $\mathcal{R} = \{\mathsf{default}, \mathsf{reverse}, \mathsf{self}\}$, as in the work of (Beck et al., 2018).

Figure 2 shows an illustrative example of a semantic graph extracted from the evidence sentences.

3.2.2 Graph Reasoning

Semantic-level graph reasoning employs a relational graph convolutional network (R-GCN) (Schlichtkrull et al., 2018) which is defined as

$$\boldsymbol{h}_i^{(l+1)} = f\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}^r_{sem}(i)}\frac{1}{|\mathcal{N}^r_{sem}(i)|}\boldsymbol{W}_r^{(l)}\boldsymbol{h}_j^{(l)} + \boldsymbol{W}_0^{(l)}\boldsymbol{h}_i^{(l)}\right)$$

where f is the *relu* activation function, $\mathcal{N}^r_{sem}(i)$ is a set of neighbors with relation r of the i-th node in $\mathcal{V}_{sem}$, and $\boldsymbol{W}_r^{(l)}, \boldsymbol{W}_0^{(l)} \in \mathbb{R}^{d_{sem}\times d_{model}}$ are weight matrices for the l-th R-GCN layer, where d_{sem} is the dimensionality of the semantic-level representation. For a word-type node $i \in \mathcal{V}_{sem}$, $\boldsymbol{h}_i^{(0)} \in \mathbb{R}^{d_{model}}$ is initialized by its span representation in the evidence sentence[5]. Finally, we ob-

[3]We use the following link of the Spacy parser: https://spacy.io/usage/linguistic-features#dependency-parse

[4]The following version of the NeuralCoref's link is used: https://github.com/huggingface/neuralcoref

[5]The span representation for a word is defined as the average pooling of the contextual representations of its all subwords. For a relation-type node $i \in \mathcal{V}_{sem}$, $\boldsymbol{h}_i^{(0)}$ is initialized by its static embedding.

tain $\boldsymbol{H}_{sem} \in \mathbb{R}^{|\mathcal{V}_{sem}| \times d_{sem}}$ as follows:

$$\boldsymbol{H}_{sem} = \boldsymbol{H}^{(L)} = \left[\boldsymbol{h}_1^{(L)}, \cdots, \boldsymbol{h}_{|\mathcal{V}_{sem}|}^{(L)}\right]$$

where L is the total number of layers used in the R-GCN for the semantic-level representation.

3.3 Semantic-infused Sentence-level Selective Graph Reasoning

In our selective graph reasoning, because there is no ground-truth answer for the nodes to be selected, we prepare K different subgraphs by applying the node selection mechanism K times, and combine the selective representations performed over K subgraphs.

3.3.1 Semantic-infused Sentence-level Representations

The first step is to obtain *semantic-infused sentence-level representations* for m evidence sentences. To this end, we construct a fully-connected *sentence-level* graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{1, \cdots, m\}$, which refers to a set of evidence sentences – $\{\mathsf{e}_1, \cdots, \mathsf{e}_m\}$. For the i-th node, we first obtain its node representation $\boldsymbol{e}'_i$ using a single feed-forward layer, as follows:

$$\boldsymbol{e}'_i = g\left(\boldsymbol{W}_{sent}\boldsymbol{E}_{i,[\mathsf{CLS}]} + \boldsymbol{b}_{sent}\right) \quad (2)$$

where g is the *gelu* activation function, and $\boldsymbol{W}_{sent}, \boldsymbol{b}_{sent}$ are the parameter weights for a linear layer. Then, for the i-th node, we further aggregate its neighbors' representations using the summation as follows:

$$\mathbf{h}'_i = \sum_{j \in \mathcal{N}_{sent}(i)} \boldsymbol{e}'_i \quad (3)$$

where $\mathcal{N}_{sent}(i)$ is a set of neighbors of the i-th node in $\mathcal{V}$.

Now, the sentence-level representation $\boldsymbol{H}_{sent} \in \mathbb{R}^{m \times d_{model}}$ is defined, as follows:

$$\boldsymbol{H}_{sent} = \left[\mathbf{h}'_1, \cdots, \mathbf{h}'_m\right] \quad (4)$$

Next, we obtain the *evidence-attentive claim representation* $\boldsymbol{h}'_{claim} \in \mathbb{R}^{d_{model}}$ and the *semantic-aware evidence representation* $\boldsymbol{H}'_{sent} \in \mathbb{R}^{m \times d_{model}}$ as follows:

$$\boldsymbol{h}'_{claim} = \mathsf{MHA}(\mathbf{C}_{[\mathsf{CLS}]}, \boldsymbol{H}_{sent}, \boldsymbol{H}_{sent}) \quad (5)$$
$$\boldsymbol{H}'_{sent} = \mathsf{MHA}(\mathbf{H}_{sent}, \boldsymbol{H}_{sem}, \boldsymbol{H}_{sem}) \quad (6)$$

where the multi-head attention (MHA) (Vaswani et al., 2017) function is defined as follows:

$$\begin{aligned} &\mathsf{MHA}\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right) = [head_1; \cdots; head_h]\boldsymbol{W}^O, \\ &head_i = \mathsf{Attn}\left(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V\right) \end{aligned} \quad (7)$$

where ; is the concatenation operator, h is the number of heads, $\boldsymbol{W}_i^Q, \boldsymbol{W}_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $\boldsymbol{W}_i^V \in \mathbb{R}^{d_{model} \times d_v}$, and $\boldsymbol{W}^O \in \mathbb{R}^{hd_v \times d_{model}}$ are weight metrices.

To combine these representations, we use the *semantic fusion* function sfu defined as:

$$\begin{aligned} \mathsf{sfu}\left(\boldsymbol{x}, \boldsymbol{y}\right) &= \boldsymbol{g} * \boldsymbol{x} + (1 - \boldsymbol{g}) * \boldsymbol{y}, \\ \boldsymbol{g} &= \sigma\big(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{W}_2\boldsymbol{y}\big) \end{aligned} \quad (8)$$

where $*$ is the element-wise operator, σ is the sigmoid function, and $\boldsymbol{W}_1, \boldsymbol{W}_2$ are weight matrices for the semantic fusion function.

Finally, the semantic-infused sentence-level representations $\boldsymbol{H}_{fused} \in \mathbb{R}^{m \times d_{model}}$ are then obtained using sfu as follows:

$$\boldsymbol{H}_{fused} = \mathsf{sfu}\left(\boldsymbol{H}'_{claim}, \boldsymbol{H}'_{sent}\right),$$

where $\boldsymbol{H}'_{claim} = [\boldsymbol{h}'_{claim}]_{i=1}^m$.

3.3.2 Node Selection Mechanism

The next step is to apply a node selection mechanism (Louis et al., 2021) that chooses a subset of nodes to be deleted[6]. First, we measure the selection probabilities $\boldsymbol{p}_{sent} \in \mathbb{R}^m$ of evidence nodes based on attention, using the claim as the query, as follows:

$$\boldsymbol{p}_{sent} = \sigma\left(g(\boldsymbol{H}_{sent}\boldsymbol{W}_3) + \boldsymbol{H}_{fused}\boldsymbol{W}_4\boldsymbol{C}_{[\mathsf{CLS}]}^T\right)$$

where $\boldsymbol{W}_3 \in \mathbb{R}^{d_{model} \times 1}$, $\boldsymbol{W}_4 \in \mathbb{R}^{d_{model} \times d_{model}}$ are weight matrices.

The node selection mechanism creates a subset of evidence nodes denoted as $\mathcal{V}'$ by filtering out the nodes with low probabilities given the threshold τ as follows:

$$\mathcal{V}' = \{j | j \in \mathcal{V} \text{ and } \boldsymbol{p}_{sent,j} \geq \tau\}$$

where $\boldsymbol{p}_{sent,j}$ is the j-th element of $\boldsymbol{p}_{sent}$. We further define $\boldsymbol{p}'_{sent} \in \mathbb{R}^m$ by zeroing the probabilities of the filtered nodes, as follows:

$$\boldsymbol{p}'_{sent} = \boldsymbol{p}_{sent} * \boldsymbol{i}_{\mathcal{V}'}$$

where $\boldsymbol{i}_{\mathcal{V}'} = [\mathcal{I}(k \in \mathcal{V}')]_{k=1}^m$ is the k-hot vector [7], and $\mathcal{I}(e)$ is the indicator function, taking the value of 1 if e is true and zero otherwise.

[6] Our node selection mechanism mostly follows the work of (Louis et al., 2021), but differs in the computation of node selection probabilities and the formula of selective aggregation.

[7] The k-hot vector has also similarly used in the work of (Cohen et al., 2019).

3.3.3 Selective Graph Reasoning

The final step is to perform selective graph reasoning using only the selected set of nodes, $\mathcal{V}'$. First, we obtain the revised fused representation $\boldsymbol{h}_i^{sel}$ for the i-th evidence sentence as follows:

$$\boldsymbol{h}_i^{sel} = \sum_{j \in N_{sent}(i)} \boldsymbol{p}'_{sent,j} \cdot \boldsymbol{H}_j^{fused}$$

Then, the reasoning-enhanced representation $\boldsymbol{h}_i^{fsel}$ is obtained as follows:

$$\begin{aligned} v_i &= \sigma\left(\langle \boldsymbol{w}_{sel}, \left[\boldsymbol{h}_i^{sel}; \boldsymbol{e}'_i\right]\rangle\right), \\ \boldsymbol{h}_i^{fsel} &= \sum_{j \in \mathcal{N}_{sent}(i)} \mathbf{p}'_{sent,j} \cdot v_j \cdot \boldsymbol{H}_j^{fused} \end{aligned}$$

where $\boldsymbol{e}'_i$ is the initial node representation defined in Eq. (2) and $\boldsymbol{w}_{sel} \in \mathbb{R}^{2d_{model}}$ is the weight vector.

We further use the residual connection to keep the initial evidence representation as follows:

$$\tilde{\boldsymbol{e}}_i = g\big(\boldsymbol{e}'_i + \mathsf{dropout}(\boldsymbol{h}_i^{fsel})\big) \tag{9}$$

where dropout is the dropout layer introduced by (Srivastava et al., 2014).

3.3.4 Ensembling Multiple Selective Graph Reasonings

Because there is no ground-truth information for nodes to be selected, we prepare multiple subgraphs by applying the node selection mechanism K times, and combine the selective reasoning-enhanced representations over K subgraphs. With the abuse of notation, suppose that $\tilde{\boldsymbol{e}}_i^{(k)}$ is the reasoning-enhanced representation of Eq. (9) yielded at the k-th selection. We take the summation of all K representations as $\sum_{k=1}^{K} \tilde{\boldsymbol{e}}_i^{(k)}$, leading to obtain $\tilde{\boldsymbol{E}}_{fsel} \in \mathbb{R}^{m \times d_{model}}$ as follows:

$$\tilde{\boldsymbol{E}}_{fsel} = \left[\sum_{k=1}^{K} \tilde{\boldsymbol{e}}_i^{(k)}\right]_{i=1}^{m} \tag{10}$$

3.4 Sequence Reasoning

Our sequence reasoning is based on MHA over *only* sentence-level evidence representations $\boldsymbol{E}_{seq} \in \mathbb{R}^{m \times d_{model}}$, described as follows.

$$\begin{aligned} \boldsymbol{E}_{seq} &= \mathsf{PE}(\boldsymbol{E}_{1,[\mathsf{CLS}]}, \cdots, \boldsymbol{E}_{m,[\mathsf{CLS}]}), \\ \boldsymbol{H}_{seq} &= \boldsymbol{E}_{seq} + \mathsf{MHA}(\boldsymbol{E}_{seq}, \boldsymbol{E}_{seq}, \boldsymbol{E}_{seq}), \end{aligned} \tag{11}$$

where PE is the absolute positional encoding (Vaswani et al., 2017).

Label	Training	Development	Test
Supported	80,035	6,666	6,666
Refuted	29,775	6,666	6,666
Not Enough Info	35,659	6,666	6,666

Table 1: Statistics of the FEVER 1.0 shared task dataset.

3.5 Prompt-based Claim Verification

Our *prompt-based claim verification* uses a task-specific template for prompt-based fine-tuning as follows: "[CLS] x_{in} It was [MASK] . [SEP]". Suppose that x_{in} is *"Roman Atwood is a content creator."*, x_{in} is converted to its prompted input "[CLS] Roman Atwood is a content creator. It was [MASK] . [SEP]". To predict [MASK], let $\mathcal{M}_{wo}\colon \mathcal{Y} \rightarrow \mathcal{V}$ be the verbalizer that converts a label into individual words. For example, $\mathcal{M}_{wo}(\mathsf{Supported})$ = "Yes", $\mathcal{M}_{wo}(\mathsf{Refutes})$ = "No", and $\mathcal{M}_{wo}(\mathsf{NotEnoughInfo})$ = "Maybe".

To determine the truthfulness of a given claim, we aggregate multiple evidence-attentive claim representations resulting from applying MHA on on $\tilde{\boldsymbol{E}}_{fsel}$ of Eq. (10) , $\boldsymbol{H}_{sem}$ in Eq. (2), and $\boldsymbol{H}_{seq}$ in Eq. (11), as follows:

$$\begin{aligned} \boldsymbol{C}_{fsel} &= \mathsf{MHA}(\mathbf{C}_{[\mathsf{CLS}]}, \tilde{\boldsymbol{E}}_{fsel}, \tilde{\boldsymbol{E}}_{fsel}), \\ \boldsymbol{C}_{sem} &= \mathsf{MHA}(\mathbf{C}_{[\mathsf{CLS}]}, \boldsymbol{H}_{sem}, \boldsymbol{H}_{sem}), \\ \boldsymbol{C}_{seq} &= \mathsf{MHA}(\mathbf{C}_{[\mathsf{CLS}]}, \boldsymbol{H}_{seq}, \boldsymbol{H}_{seq}), \\ \mathbf{H} &= \boldsymbol{W}_{claim}([\boldsymbol{C}_{fsel}; \boldsymbol{C}_{sem}; \boldsymbol{C}_{seq}]), \end{aligned} \tag{12}$$

where $W_{claim} \in \mathbb{R}^{d_{model} \times 3d_{model}}$ is a trainable parameter matrix.

Given a claim-evidence example (c, e), where $\mathsf{e} = (\mathsf{e}_1, \cdots, \mathsf{e}_m)$, the probability of label y is computed as follows:

$$\begin{aligned} \mathsf{p}(\mathsf{y}|\mathsf{c},\mathsf{e}) &= \mathsf{p}\big([\mathsf{MASK}] = \mathcal{M}_{wo}(\mathsf{y})|\mathsf{c},\mathsf{e}\big) \\ &= \frac{\exp\big(\boldsymbol{w}_{\mathcal{M}_{wo}(\mathsf{y})} \boldsymbol{H}_{[\mathsf{MASK}]}\big)}{\sum_{\mathsf{y}' \in \mathcal{Y}} \exp\big(\boldsymbol{w}_{\mathcal{M}_{wo}(\mathsf{y}')} \boldsymbol{H}_{[\mathsf{MASK}]}\big)}, \end{aligned} \tag{13}$$

where $\boldsymbol{w}_{\mathcal{M}_{wo}}(\mathsf{y})$ is the output embedding for the label word of $\mathcal{M}_{wo}(\mathsf{y})$ for y, and $\boldsymbol{H}_{[\mathsf{MASK}]}$ is the contextual representation [MASK] token in $\boldsymbol{H}$.

4 Experiments

4.1 Experimental Setting

Dataset

We used FEVER, which is a large-scale public dataset, for fact verification. (Thorne et al.,

Model	Dev		Test	
	LA	F.S	LA	F.S
UNC NLP	69.72	66.49	68.21	64.21
GEAR (BERT$_{\text{base}}$)	74.84	70.69	71.60	67.10
DREAM (XLNet$_{\text{large}}$)	79.16	-	76.85	70.60
KGAT (BERT$_{\text{large}}$)	77.91	75.86	73.61	70.24
⌞ (RoBERTa$_{\text{large}}$)	78.29	76.11	74.07	70.38
LOREN (BERT$_{\text{large}}$)	78.44	76.21	74.43	70.71
⌞ (RoBERTa$_{\text{large}}$)	81.14	78.83	76.42	72.93
MLA (RoBERTa$_{\text{large}}$)	79.31	75.96	77.05	73.72
Ours (RoBERTa$_{\text{large}}$)	**83.13**	**79.87**	**77.50**	**73.90**

Table 2: Fact verification results on the dev and blind test set of FEVER task, where F.S (FEVER score) is the main evaluation metric. The best is **bolded** text, and the second best is underlined.

Model	Dev		Test	
	LA	F.S	LA	F.S
MLA	79.31	75.96	77.05	73.72
SISER⋆	83.13	79.85	76.82	73.18
SISER∘ ($\tau = 0.49$)	82.62	79.40	77.18	73.48
SISER ($\tau = 0.49$)	83.13	**79.87**	**77.50**	**73.90**

Table 3: Ablation study for the semantic-infused sentence-level selective graph reasoning and the sequence reasoning on FEVER development and blind test set. ⋆ and ∘ denote the run without the semantic-infused sentence-level selective graph reasoning and the sequence reasoning, respectively.

2018a,b), which was split into *training*, *development*, and *blind test* set in our experiments. FEVER consists of 185,455 annotated claims with 5,416,537 Wikipedia documents, where claims are classified as *Supported*, *Refuted*, or *Not Enough Info*. Because we use prompt-based fine-tuning, all labels are verbalized as *Yes*, *No*, or *Maybe*. Table 1 shows more detailed statistics for the FEVER dataset. The performance of the evidence sentence retrieval methods are presented in Appendix B.

Evaluation Metrics

The official evaluation metrics are Label Accuracy (LA) and FEVER Score (F.S)[8]. Label Accuracy is a general evaluation metric, which is the accuracy of the predicted label for a claim regardless of the retrieved evidence.

4.2 Main Results

The fact verification performance is presented in Table 2. In the large-size PLM settings, SISER outperforms the best baseline model by increasing Label Accuracy and FEVER Score by 0.45 and 0.18, respectively.

For a fair comparison, we also compare SISER with KGAT and LOREN, which employ the same setting of using PLM and evidence retrieval, while MLA, the state-of-the-art baseline model, is different from ours in using evidence retrieval. As shown in Table 2, SISER outperforms KGAT and LOREN, which employ only sentence-level interaction among evidences. The results may support our motivation that the combination of the three types of reasoning (i.e., semantic-level graph-reasoning, semantic-infused sentence-level selective graph-reasoning, and sequence reasoning) is helpful to address the aformentioned 'unit-biased reasoning' and 'oversmoothing' problems of the existing graph-based approaches.

Model	Dev		Test	
	LA	F.S	LA	F.S
$\tau = 0.0^{\bullet}$	83.07	79.84	77.07	73.65
$\tau = 0.35$	83.00	79.74	77.11	73.70
$\tau = 0.40$	83.05	79.84	77.00	73.63
$\tau = 0.45$	82.98	79.69	76.86	73.66
$\tau = 0.49$	**83.13**	**79.87**	**77.50**	**73.90**
$\tau = 0.60$	83.04	79.80	77.30	73.68

Table 4: Ablation study of the node selection mechanism for varying values of the node masking rate τ. • denotes the fully-connected setting.

Model	Dev		Test	
	LA	F.S	LA	F.S
SISER⋆	83.05	79.77	76.82	73.18
SISER	**83.13**	**79.87**	**77.50**	**73.90**

Table 5: Ablation study for the prompt-based learning vs. the conventional fine-tuning on the FEVER development set. ⋆ denotes the conventional fine-tuning.

Model	Dev		Test	
	LA	F.S	LA	F.S
MLA	79.31	75.96	77.05	73.72
SISER⋆ ($\tau = 0.49$)	79.88	75.04	**77.96**	73.06
SISER ($\tau = 0.49$)	**83.13**	**79.87**	77.50	**73.90**

Table 6: Ablation study for examining the effect of evidence retrieval. ⋆ denotes the run based on the evidence retrieval of MLA (Kruengkrai et al., 2021).

[8]https://github.com/sheffieldnlp/fever-scorer

(a)	**Claim:**	Liam Neeson has been nominated for a British Academy of Film and Television Arts award.
	Evidence:	[Liam Neeson] (12-th sentence in wiki page) He has been nominated for a number of awards, including an Academy Award for Best Actor, a BAFTA Award for Best Actor in a Leading Role and three Golden Globe Awards for Best Actor in a Motion Picture Drama.
	Label:	SUPPORTS **Predicted Label:** NOT ENOUGH INFO
(b)	**Claim:**	LinkedIn is limited to 24 languages as of 2015.
	Evidence:	[LinkedIn] (15-th sentence in wiki page) Based in the United States, the site is, as of 2013, available in 24 languages, including Arabic, Chinese, English, French, German, Italian, Portuguese, Spanish, Dutch, Swedish, Danish, Romanian, Russian, Turkish, Japanese, Czech, Polish, Korean, Indonesian, Malay, and Tagalog.
	Label:	SUPPORTS **Predicted Label:** REFUTES
(c)	**Claim:**	SZA is an American Neo Soul singer.
	Evidence:	[SZA (singer)] (7-th sentence in wiki page) SZA is a Neo Soul singer whose music is described as Alternative RB , with elements of soul , hip hop , minimalist RB , cloud rap , ethereal RB , witch house and chillwave. [SZA (singer)] (1-th sentence in wiki page) Solána Rowe (born November 8, 1990), better known by her stage name SZA, is an American singer songwriter.
	Label:	SUPPORTS **Predicted Label:** SUPPORTS

Figure 3: Error analysis of SISER: (a) and (c): the cases of requiring more elaborated and mulit-hop reasoning; (b): the case of a human annotation error.

4.3 Ablation Study

The Effect of Using Semantic-infused Sentence-level Selective Graph Reasoning

To evaluate the effect of using semantic-infused sentence-level selective graph reasoning in Section 3.3, Table 3 shows the comparison results of SISER with and without semantic-infused sentence-level selective graph reasoning on the FEVER development and blind test sets. It is shown that the use of semantic-infused selective graph reasoning leads to improved performance in terms of both Label Accuracy and the FEVER Score.

It is remarkable that SISER$\star$, even without using semantic-infused selective graph reasoning, outperforms MLA in the development set. While (Kruengkrai et al., 2021) argued that graph reasoning may not be necessary, given the improved performance of the MLA, our results indicate that this argument is still controversial, and suggest that graph reasoning has the potential to make further improvements and needs to be explored for fact verification while carefully avoiding the limitations of GNNs.

The Effect of Using Sequence Reasoning

Table 3 further presents the performance of SISER when sequence reasoning is excluded (referred to as SISER$\circ$), that is, without using C_{fsel} in Eq (12)). As shown in Table 3, SISER$\circ$ leads to improvements over LOREA, indicating that the performance achieved by SISER in Table 2 is not obtained simply by incorporating sequence reasoning but dominantly by equipping with the proposed manner of graph reasoning. In particular, SISER$\circ$ shows an increases in Label Accuracy by approximately 1.5 over LOREN on the development set, whereas SISER with sequence reasoning demonstrates only a slight increase of approximately 0.5 in Label Accuracy. A similar tendency is observed in the blind test set; SISER$\circ$ makes the increase of 0.76 in Label Accuracy over LOREN, which is larger than the increase of 0.32 obtained by SISER with sequence reasoning.

The Effect of Choosing Evidence Retrieval

In Table 2, while SISER shows consistent improvements over MLA on the development and test sets, a significant difference in performance gains is noticeable between the two sets. SISER achieves a large performance gain over MLA on the development set, increasing the Label Accuracy and FEVER Score by 3.82 and 3.91, respectively, while only a slight improvement on the blind test set is observed, exhibiting an increase of 0.45 in Label Accuracy and 0.18 in FEVER Score.

We believe that the main reason for this discrepancy between development and test sets results from the different evidence retrieval methods between SISER and MLA, i.e., while SISER and LOREN adopt KGAT's evidence retrieval, MLA uses its own evidence retrieval. In particular, the retrieval performances of the top 5 evidence sentences resulting from MLA and KGAT are substantially changed between the development and test sets, as shown in Table 7. In terms of Recall@5, the retrieval performances on the "development set" are largely different between KGAT and MLA (i.e., 94.57 for KGAT and 88.64 for MLA), whereas the retrieval performances on the "test set" of both methods are fairly similar (i.e., 87.47 for KGAT and 87.58 for MLA). Given this observation, the substantially improved performance of SISER over MLA on the development set (Table 2) may primarily originate from the large recall performance of the evidence retrieval of KGAT, and not from the proposed enhanced graph reasoning components.

For a fair comparison with MLA, Table 6 presents the results of SISER based on MLA's evidence retrieval (SISER$\star$). In terms of on FEVER Score, SISER$\star$ does not lead to improvements over MLA, even exhibiting performance degradation, in contrast to the SISER that uses KGAT's retrieval. Nevertheless, SISER$\star$ leads to further improvements over MLA in Label Accuracy, particularly in achieving a state-of-the-art performance on the blind test set.

As MLA is considered as an advanced approach to sequence reasoning without relying on graph reasoning, we believe that the enhanced graph reasoning modules in SISER are 'complementary' to MLA for further improvement; for example, including a simple combination by using MLA as an alternative module of sequence reasoning in SISER.

Evaluation of Node Selection Mechanism

To examine the effect of the node selection mechanism in Section 3.3.2, Table 4 shows the comparison results of SISER with varying values of τ. It is shown that $\tau = 0.49$ outperforms the fully-connected setting ($\tau = 0.0$). The results imply that the node selection mechanism based on the selection probabilities may be helpful in obtaining irrelevance-free evidence representations, related to the oversmoothing issue of GNNs.

Prompt-based Learning versus Conventional Fine-tuning

To examine the effect of prompt-based claim verification, Table 5 compares the results of SISER when using prompt-based learning or conventional fine-tuning. It is clearly shown that the use of prompt-based learning outperforms conventional fine-tuning, likely reducing the gap between the tasks used in pre-training and the fine-tuning.

4.4 Case Study

As shown in Figure 3, we present three examples for analyzing the prediction errors of SISER.

In Figure 3 (a), the SISER prediction for this case is "*Not Enough Info*". From our analysis, this case requires the complex reasoning ability to understand "*a BAFTA award*," which is the abbreviation of "*a British Academy of Film and Television Arts award*". However, in Figure 3 (c), the case requires multi-hop complex reasoning to predict the claim; the claim "*SZA is an **American Neo Soul singer***" is supported by multiple pieces of evidence sentences.

In Figure 3 (b), it seems that this case originates from a human annotation error, as also discussed by (Kruengkrai et al., 2021). The claim "*LinkedIn is limited to 24 languages as of 2015*" is not supported by evidence.

5 Conclusion

In this paper, we propose SISER for fact verification, which combines three types of reasoning (i.e., semantic-level graph reasoning, semantic-infused sentence-level selective graph reasoning, and sequence reasoning) by addressing two potential limitations of graph reasoning — the "unit-biased reasoning" and the "over-smoothing" problems. The experimental results obtained using the FEVER dataset showed that the proposed SISER outperformed other graph-based approaches and achieved state-of-the-art performances in both the development and test sets.

In future work, we would like to incorporate semantic-level and semantic-fused graph reasoning into evidence retrieval and explore the joint learning framework of evidence retrieval and claim verification in a multi-task learning setting.

Acknowledgements

This work was supported by NAVER Corp. We would like to thank all anonymous reviewers for their valuable comments and suggestions.

References

Gabor Angeli and Christopher D. Manning. 2014. NaturalLI: Natural logic inference for common sense reasoning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 534–545, Doha, Qatar. Association for Computational Linguistics.

Daniel Beck, Gholamreza Haffari, and Trevor Cohn. 2018. Graph-to-sequence learning using gated graph neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 273–283, Melbourne, Australia. Association for Computational Linguistics.

Deli Chen, Yankai Lin, Wei Li, Peng Li, Jie Zhou, and Xu Sun. 2020a. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3438–3445.

Jiangjie Chen, Qiaoben Bao, Changzhi Sun, Xinbo Zhang, Jiaze Chen, Hao Zhou, Yanghua Xiao, and

Lei Li. 2020b. Loren: Logic-regularized reasoning for interpretable fact verification.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Liying Cheng, Dekun Wu, Lidong Bing, Yan Zhang, Zhanming Jie, Wei Lu, and Luo Si. 2020. ENT-DESC: Entity description generation by exploring knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1187–1197, Online. Association for Computational Linguistics.

William W. Cohen, Haitian Sun, R. Alex Hofer, and Matthew Siegler. 2019. Differentiable representations for multihop inference rules. *CoRR*, abs/1905.10417.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ning Ding, Yulin Chen, Xu Han, Guangwei Xu, Pengjun Xie, Hai-Tao Zheng, Zhiyuan Liu, Juanzi Li, and Hong-Gee Kim. 2021. Prompt-learning for fine-grained entity typing. *CoRR*, abs/2108.10604.

Matthias Fey and Jan E. Lenssen. 2019. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. 2019. Predict then propagate: Graph neural networks meet personalized pagerank. In *International Conference on Learning Representations (ICLR)*.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.

Christopher Hidey and Mona Diab. 2018. Team SWEEPer: Joint sentence extraction and fact checking with pointer networks. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 150–155, Brussels, Belgium. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1).

Xiaomeng Hu, Shi Yu, Chenyan Xiong, Zhenghao Liu, Zhiyuan Liu, and Ge Yu. 2022. P^3 ranker: Mitigating the gaps between pre-training and ranking fine-tuning with prompt-based learning and pre-finetuning. *CoRR*, abs/2205.01886.

Yinya Huang, Meng Fang, Yu Cao, Liwei Wang, and Xiaodan Liang. 2021. DAGN: Discourse-aware graph network for logical reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5848–5855, Online. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Canasai Kruengkrai, Junichi Yamagishi, and Xin Wang. 2021. A multi-level attention model for evidence-based fact checking. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2447–2460, Online. Association for Computational Linguistics.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. AAAI Press.

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. *CoRR*, abs/1907.11692.

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*, pages 7342–7351, Online. Association for Computational Linguistics.

Steph-Yves M. Louis, Alireza Nasiri, Fatima J. Rolland, Cameron Mitro, and Jianjun Hu. 2021. NODE-SELECT: A graph neural network based on A selective propagation technique. *CoRR*, abs/2102.08588.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6859–6866. AAAI Press.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. 2020. Dropedge: Towards deep graph convolutional networks on node classification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

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

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4603–4611. PMLR.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. 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)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *International Conference on Learning Representations*. Accepted as poster.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Lingxiao Zhao and Leman Akoglu. 2020. Pairnorm: Tackling oversmoothing in {gnn}s. In *International Conference on Learning Representations*.

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*, pages 6170–6180, Online. Association for Computational Linguistics.

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*, pages 892–901, Florence, Italy. Association for Computational Linguistics.

A Implementation Details

SISER was implemented by using PyTorch (Paszke et al., 2019) and HuggingFace Transformers (Wolf et al., 2020). Additionally, the PyTorch-Geometric and SpaCy (Fey and Lenssen, 2019; Honnibal and Montani, 2017) were used for graph modeling and dependency parsing. Experiments were conducted using 4 Nvidia RTX A6000 GPU. All optimizations were performed using the Adafactor optimizer (Shazeer and Stern, 2018) with a linear warm-up of the learning rate. The warmup proportion was 0.06. The batch size and accumulation steps were 8 and 8, respectively. That is, the total batch size is 256. Gradients were clipped if their norms exceeded 1.0. The number of K sub-graphs was 6 and $\tau = 0.49$. In supervised learning, our loss $\mathcal{L}$ can be fine-tuned to minimize the weighted cross-entropy loss introduced by MLA (Kruengkrai et al., 2021).

Our hyperparameter is summarized as below:

- Optimizer: Adafactor
- Learning rate: $2e-5$
- warmup proportion: 0.06
- Number of sub-graph: 6
- Total Batch size: 256
- Gradient norm: 1.0
- Node masking rate τ: 0.49
- Label words: $\mathsf{Supported}$: Yes, $\mathsf{Refuted}$: No, $\mathsf{Not\ Enough\ Info}$: Maybe

Data	Method	Prec@5	Recall@5	F1@5
Dev	UNC NLP*	36.49	86.79	51.38
	GEAR*	**40.60**	86.36	**55.23**
	KGAT$^\diamond$	27.29	**94.37**	42.34
	DREAM$^\diamond$	26.67	87.64	40.90
	MLA$^\diamond$	25.63	88.64	39.76
	monoT5$^\bullet$	25.66	90.54	37.17
Test	KGAT$^\diamond$	25.21	87.47	39.14
	MLA$^\diamond$	**25.33**	**87.58**	**39.29**

Table 7: Results of the sentence selection methods in the precision@5, recall@5, and F1@5 metrics on the FEVER development set and blind test set, respectively. $*, \diamond, \bullet$ denote ESIM-based retrieval model, BERT-based retrieval model, and T5-base model, respectively.

B Evidence Sentence Retrieval

Since our work focuses on claim verification, we directly adapt the evidence retrieval method from KGAT (Liu et al., 2020). As shown in Table 7, KGAT shows the best Recall@5 performance for sentence selection on the FEVER development set. Different from the result on the FEVER development set, MLA shows the better Recall@5 performance than KGAT.