Denoising Rationalization for Multi-hop Fact Verification via Multi-granular Explainer

Jiasheng Si[1,4†], **Yingjie Zhu**[2,3†], **Wenpeng Lu**[1,4], **Deyu Zhou**[2,3*]
[1]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences)
[2] School of Computer Science and Engineering, Southeast University, China
[3] Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
[4] Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, China
{jiashengsi, lwp}@qlu.edu.cn, zhu359107@gmail.com, d.zhou@seu.edu.cn

Abstract

The success of deep learning models on multi-hop fact verification has prompted researchers to understand the behavior behind their veracity. One feasible way is erasure search: obtaining the rationale by entirely removing a subset of input without compromising verification accuracy. Despite extensive exploration, current rationalization methods struggle to discern nuanced composition within the correlated evidence, which inevitably leads to noise rationalization in multi-hop scenarios. To address this issue, this paper proposes the consistent multi-granular rationale extraction method, aiming to realize the denoising rationalization for multi-hop fact verification. Specifically, given a pre-trained veracity prediction model, two independent external explainers are introduced and trained collaboratively to enhance the discriminating ability by imposing varied constraints. Meanwhile, three key properties (*Fidelity*, *Consistency*, *Salience*) are introduced to regularize the denoising and faithful rationalization process. Additionally, a new *Noiselessness* metric is proposed to measure the purity of the rationales. Experimental results on three multi-hop fact verification datasets show that the proposed approach outperforms 12 baselines.

1 Introduction

Computational multi-hop fact verification approaches typically explore neural models to verify the truthfulness of claims through multi-hop reasoning across multiple pieces of evidence (Jiang et al., 2020; Ostrowski et al., 2021). Despite the prevalence of these approaches, limited attention has been given to elucidating the underlying rationale of these systems(Kotonya and Toni, 2020a; Si et al., 2023), which compromises user trust in the prediction and hinders transparency of the models (Lyu et al., 2022; Tang et al., 2021).

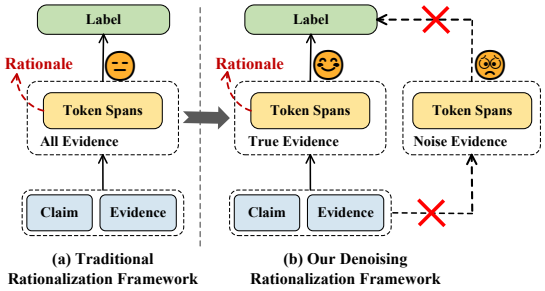

Figure 1: A comparison between traditional and denoising rationalization framework.

A clear-cut rationalization way to explain the behavior of a model regarding a specific prediction is **erasure search** (Li et al., 2016; Feng et al., 2018; De Cao et al., 2020), a post-hoc approache wherein the *rationale* is derived by searching for a maximum subset of the input (e.g., token span) that can be entirely removed without impacting the model prediction. This removal perturbation to the input guarantees the explicit decorrelation of discarded features with model prediction, thereby showing faithfulness to the model for the derived rationales (Si et al., 2023). In this paper, we explore a novel denoising rationalization paradigm to extract potential post-hoc rationales for explainable multi-hop fact verification.

Multi-hop fact verification entails a complex reasoning scenario with distinct constituent elements as input (i.e., *claim*, *true evidence*, and *noise evidence*.)[1](Si et al., 2023). A reasonable multi-hop fact verification model ought to be capable of discerning this discrepancy by aggregating information solely from the true evidence to reach its prediction, while disregarding the noise evidence (i.e.,

[†] Equal Contribution.
[*] Corresponding Author.

[1]*True evidence* provides intrinsic multi-hop information capable of truly verifying the claim. *Noise evidence* constitutes extracted evidence from the web containing highly semantic and linguistic content related to the claim, but is irrelevant to the claim verification.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 12593–12608

Right Prediction for Right Reasons (Gupta et al., 2022)). However, as depicted in Fig.1(a), current rationalization methods follow a paradigm stemming from the explainability research on text classification tasks, which typically extract post-hoc rationales (token span) by regarding different types of evidence as **equally** (Li et al., 2016; Paranjape et al., 2020; Wiegreffe and Marasovic, 2021). In this case, a consequential issue arises: extensive confusing noise token spans will inevitably be extracted as rationales from the noise evidence, i.e., *noise rationalization*, due to the lack of discriminating ability about the nuanced composition within the correlated evidence. This noise rationalization implies the containing of irrelevant and unfaithful rationales to the task prediction. As shown in Fig.2, VMASK (Chen and Ji, 2020) extract token rationales from both true evidence (token spans in *E1*, *E3*, and *E4*) and noise evidence (token spans in *E2*, *E5*) without differentiation, while the token spans in *E2* and *E5* are unexpected to affect the reasoning process of the model (i.e., *Right Prediction for Wrong Reasons*). In essence, this rationalization uses spurious correlations between irrelevant evidence, the claim, and the inference label to extract rationales.

This paper argues that a rationalization system for multi-hop fact verification is considered reasonable only when it can extract sufficient, concise, and pure rationale to reflect the basis of its decisions, which is to extract the *"right tokens"* from the *"right sentences"*, as illustrated in Fig.1(b). For example, in Fig.2, an ideal rationalization system could solely retain the the task-relevant rationales in {*E1*, *E3*, *E4*} and wholly eliminate the irrelevant rationales in {*E2*, *E5*}. To tackle this issue, we propose a **C**onsistent m**U**lti-granular **R**ationale **E**xtraction (CURE) approach for denoising rationalization in multi-hop fact verification. The **core idea** of our work is that we introduce two independent explainers in parallel, which enables token and sentence explainers to assist and inhibit each other collaboratively by imposing varied constraints between them. This mutual effect enhances the discriminating ability of the token/sentence explainer by taking the sentence/token explainer as intermediate, thus amplifying true tokens from true evidence and suppressing noise tokens from noise evidence.

In specific, given a pre-trained multi-hop fact verification model, two parameterized explainers are first trained to generate mask vectors for each token and sentence to indicate which token or sentence is necessary or can be discarded. Then, the valid rationales are derived by intersecting the two granular mask vectors and applying them to perturb the input. Meanwhile, inspired by (Si et al., 2021), the sentence mask vector is used to intervene in the coupling coefficients during the evidence aggregating process. Furthermore, three properties are introduced to regularize *denoising* and *faithful* rationalization process: (i) *Fidelity* to constrain the faithfulness of rationales; (ii) *Consistency* to increase the consistency across the multi-granular explainers; (iii) *Salience* to guide the rationale extraction through a predefined salience score. Additionally, a new metrics *Noiselessness* is proposed to evaluate the purity of the extracted rationales. We empirically conduct the experiments on 3 different multi-hop fact verification datasets over 12 baselines. Both the automatic and manual evaluation results validate the superiority of our method.

Claim
This organism and Panax are both plant genera. The Gulf named after the organism is part of Port McArthur Tidal Wetlands System.

Evidence
E_1(✓): [Port McArthur Tidal Wetlands System] The Port McArthur Tidal Wetlands System comprises a 994 km tract of tidal wetlands on the south-west coast of the Gulf of Carpentaria in the Northern Territory of Australia.
E_2(✗): [Port McArthur Tidal Wetlands System] The land extends along the coast opposite the Sir Edward Pellew Group of Islands, incorporating the estuaries of the McArthur and Wearyan Rivers.
E_3(✓): [Panax] The Panax (ginseng) genus belongs to the Araliaceae (ivy) family.
E_4(✓): [Carpentaria] Carpentaria acuminata (carpentaria palm), the sole species in the genus Carpentaria, is a palm native to tropical coastal regions in the north of Northern Territory, Australia.
E_5(✗): [Sir Edward Pellew Group of Islands] The Sir Edward Pellew Group of Islands is situated in the south-west corner of the Gulf of Carpentaria, off the northern coast of Australia.

Label SUPPORTS

✓ ✗ : Ground-truth Sentence Rationale Label

Figure 2: An example in the HoVer dataset (Jiang et al., 2020) with marked token rationales extracted by VMASK (Chen and Ji, 2020).

2 Preliminaries

Task. Given a claim $\boldsymbol{c}$ with associated evidence $E = \{\boldsymbol{e_1}, \boldsymbol{e_2}, \ldots, \boldsymbol{e_n}\}$, we construct a fully connected input graph $G = (X, A)$, where n is the number of evidence, $\boldsymbol{x_i} \in X$ denotes i-th evidence node by concatenating the evidence text $\boldsymbol{e_i}$ with the claim $\boldsymbol{c}$. We first pretrained a Veracity Prediction model (Si et al., 2021) to verify the claim, then two independent explainers are trained jointly to extract the rationales with multiple granularity, i.e., sentence rationales $R_s = (X_s \subset X, A_s \in \mathbb{R}^{n \times n})$ and token rationales $\boldsymbol{r} = \{\boldsymbol{r}_i \subset \boldsymbol{e}_i\}|_{i=0}^{n}$, where

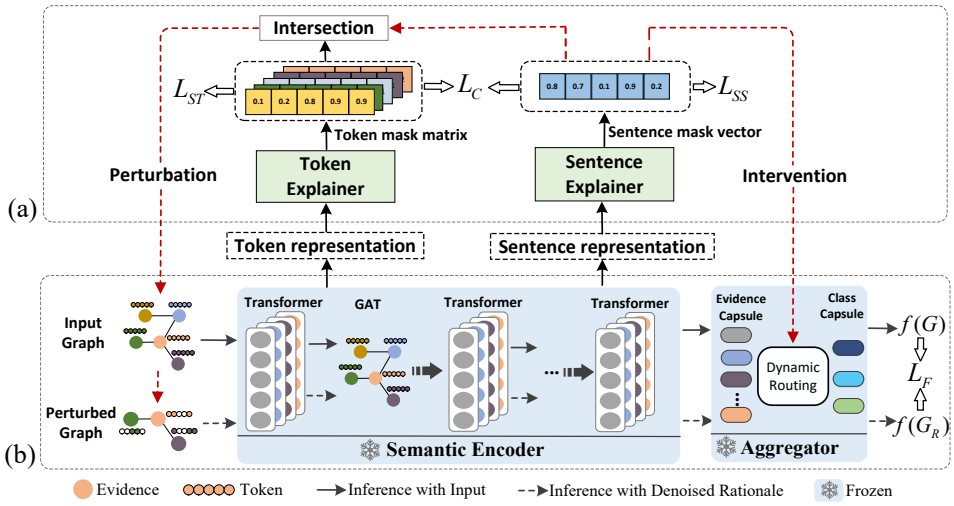

Figure 3: The overall architecture of our CURE. (a) Multi-granular rationale extraction with two parallel explainers to yield token and sentence mask vectors. (b) Veracity prediction model with parameters fixed to verify the claim.

$\boldsymbol{r}_i = \{\boldsymbol{t}_{i,j} | \boldsymbol{t}_{i,j} \in \boldsymbol{e}_i\}$. Finally, the valid multi-granular rationale G_R is obtained.

Definition. Following Jain et al. (2020), we define the *faithfulness* and *Noiselessness* in a multi-granular rationale extraction scenario.

Definition 1. *(Faithfulness) R_s and $\boldsymbol{r}$ are multi-granular faithful to their corresponding prediction Y if and only if Y rely entirely on $G_R = (\{\boldsymbol{x}_i \cap \boldsymbol{r}_i \mid \boldsymbol{x}_i \in X_s\}, A_s)$.*

Definition 2. *(Noiselessness) R_s and $\boldsymbol{r}$ are multi-granular noiseless if and only if satisfying*

$$\sum_{\boldsymbol{x_i} \in X_s} |\boldsymbol{x_i} \cap \boldsymbol{r}_i| \leq \epsilon, \sum_{\boldsymbol{x'_i} \in X \setminus X_s} |\boldsymbol{x'_i} \cap \boldsymbol{r}_i| \to 0, \quad (1)$$

where ϵ is the maximum expected sparsity of token rationales. $X \setminus X_s$ denotes the complementary subset of X_s.

3 Method

As shown in Fig.3, we propose the method **CURE**, which includes: (i) the veracity prediction module (Fig.3(b)), (ii) multi-granular rationale extraction module (Fig.3(a)), and the terms we optimize: (iii) the key properties, (iv) the optimization.

3.1 Veracity Prediction

For the multi-hop fact verification model $f(\cdot)$, as shown in Fig.3(b), we adopt the typical veracity prediction model as illustrated in Si et al. (2021).

(I) Semantic Encoder Given the input graph $G = (X, A)$, a Transformer (Vaswani et al., 2017) layer is first applied to the node X to obtain the token representation $\boldsymbol{h} = \langle \boldsymbol{h}_0, \boldsymbol{h}_1, ..., \boldsymbol{h}_n \rangle$ for each evidence, where $\boldsymbol{h}_i = \langle \boldsymbol{h}_{i,0}, \boldsymbol{h}_{i,1}, ..., \boldsymbol{h}_{i,|\boldsymbol{x}_i|} \rangle$, $\boldsymbol{h}_{i,j \neq 0}$ denotes the j-th **token representation** in i-th evidence. Then, a GAT layer is applied to the *[CLS]* token representation to propagate the message exchange among all evidence along the edges, i.e., $\tilde{\boldsymbol{h}}_{i,0}|_{i=0}^{n} = GAT(\boldsymbol{h}_{i,0}|_{i=0}^{n})$. Thus, the updated representation is obtained, $\boldsymbol{h}_i = \langle \tilde{\boldsymbol{h}}_{i,0}, \boldsymbol{h}_{i,1}, ..., \boldsymbol{h}_{i,|\boldsymbol{x}_i|} \rangle$, where $\tilde{\boldsymbol{h}}_{i,0}$ denotes the **sentence representation** for i-th evidence. By stacking L-layers of Transformer with GAT, we get the representation $\boldsymbol{H} = \langle \boldsymbol{h}^0, \boldsymbol{h}^1, ..., \boldsymbol{h}^L \rangle$, where $\boldsymbol{h}^0 = X$ and the l layer $\boldsymbol{h}^l = \{\boldsymbol{h}_i\}|_{i=0}^{n}$.

(II) Evidence Aggregator To reach the claim verification based on the evidence, a capsule network (Sabour et al., 2017) is used to aggregate the information among all the evidence by taking sentence representation $\tilde{\boldsymbol{h}}_{i,0}^{L}|_{i=0}^{n}$ as the evidence capsule and label as the class capsule. It permits us to further eliminate the effect of non-rationale for veracity prediction.

The veracity prediction model $f(\cdot)$ is optimized by the capsule loss, and the parameters are **frozen** during the subsequent rationalization process.

3.2 Multi-granular Rationale Extraction

Mask Learning To realize the valid multi-granular perturbation to the input graph G, as shown in Fig.3(a), our CURE relies on two independent parameterized explainers to generate binary masks at the token level and the sentence level, indicating the absence or presence of each token or sentence.

In specific, taking the hidden representation $\boldsymbol{H}$ in the Semantic Encoder as input, for the token-level explainer, we employ a shallow interpreter

network $g_t(\cdot)$ (i.e., one-hidden-layer MLP network (De Cao et al., 2020)) to yield binary token mask matrix $\boldsymbol{z} = \{\boldsymbol{z}_i\}|_{i=0}^{n}$ based on the token representation, where $\boldsymbol{z}_i = \{z_{i,j}\}|_{j=1}^{|\boldsymbol{x}_i|}$ denotes mask vector for i-th evidence, $|x_i|$ is the number of tokens. The sentence representation with $j = 0$ is not considered here. We then apply Hard Concrete reparameterization (HCR) trick (Louizos et al., 2018) to enforce the values approximate to discrete 0 or 1, while keeping continuous and differential for learning token level mask matrix.

$$\begin{aligned} \boldsymbol{z}_i = \boldsymbol{z}_i^0 \odot \cdots \odot \boldsymbol{z}_i^L, \quad \boldsymbol{pt}_i = \boldsymbol{pt}_i^0 \odot \cdots \odot \boldsymbol{pt}_i^L, \\ (\boldsymbol{z}_i^l, \boldsymbol{pt}_i^l) = \text{HCR}(g_t(\boldsymbol{h}_{i,j}^l|_{j=1}^{|\boldsymbol{x}_i|})), \end{aligned} \tag{2}$$

where $\odot$ is Hadamard product, $pt_{i,j} \in \boldsymbol{pt}_i$ is the importance score of j-th token in i-th evidence.

For the sentence-level explainer, we train a separate interpreter network $g_s(\cdot)$ to predict a binary sentence mask vector $\boldsymbol{m} \in \mathbb{R}^n$ based on sentence representation to indicate the absence of sentence,

$$\begin{aligned} \boldsymbol{m} = \boldsymbol{m}^0 \odot \cdots \odot \boldsymbol{m}^L, \quad \boldsymbol{ps} = \boldsymbol{ps}^0 \odot \cdots \odot \boldsymbol{ps}^L, \\ (\boldsymbol{m}^l, \boldsymbol{ps}^l) = \text{HCR}(g_s(\tilde{\boldsymbol{h}}_{i,0}^l|_{i=0}^{n})). \end{aligned} \tag{3}$$

Input Perturbation Unlike previous work with single granular rationales (De Cao et al., 2020), to obtain the denoising rationales, we need to collaboratively aggregate the results of two explainers to perturb the input G. In addition, inspired by Si et al. (2023), our CURE capture the topological relationship between different evidence by re-construct the adjacency matrix A_s. Specifically, the valid rationale G_R can be derived by intersecting the two granular masks on the original input graph,

$$\begin{aligned} \boldsymbol{r} = \{\boldsymbol{r}_i\}|_{i=0}^{n}, \text{where } \boldsymbol{r}_i = \boldsymbol{x}_i \odot \boldsymbol{z}_i, \\ R_s = (X_s = X \odot \boldsymbol{m}, A_s = A \odot \boldsymbol{m}^\top \boldsymbol{m}), \\ G_R = (\{\boldsymbol{x_i} \cap \boldsymbol{r}_i \mid \boldsymbol{x_i} \in X_s\}, A_s). \end{aligned} \tag{4}$$

Notably, to ensure that only extracted rationales are used for veracity prediction, we further intervene in the dynamic routing of the capsule network for succinct aggregation by multiplying the sentence mask vector with the coupling coefficients.

3.3 Properties

Fidelity Fidelity guarantees that the model veracity is maintained after perturbing the input, which measures the sufficiency for *faithfulness* of multi-granular rationales (Jiang et al., 2021). To ensure the faithfulness of rationales, We re-feed the original graph G and perturbed graph G_R into the veracity model $f(\cdot)$ to generate the prediction logits, respectively. Then the Euclidean distance between these two logits is defined as fidelity loss,

$$\mathcal{L}_F = \|f(G) - f(G_R)\|_2. \tag{5}$$

Consistency We derive the denoised token rationale via improving the *consistency* between the two independent explainers (i.e., $g_t(\cdot)$ and $g_s(\cdot)$), which ensures the synergy through the interplay between them. We introduce the symmetric Jensen-Shannon Divergence to regularize the consistency between the importance score of two mask vectors,

$$\begin{aligned} \mathcal{L}_C = &\frac{1}{2}\text{KL}(P(\boldsymbol{z})||\frac{P(\boldsymbol{z}) + P(\boldsymbol{m})}{2}) \\ &+ \frac{1}{2}\text{KL}(P(\boldsymbol{m})||\frac{P(\boldsymbol{z}) + P(\boldsymbol{m})}{2}), \end{aligned} \tag{6}$$

where $P(\boldsymbol{z}) = \text{softmax}_i(\sum_{j=1}^{|\boldsymbol{x}_i|} pt_{i,j})$, $P(\boldsymbol{m}) = \text{softmax}_i(ps_i)$, and $\text{KL}(\cdot||\cdot)$ denotes the Kullback-Leibler divergence.

Salience (I) Salience-Sentence: The sentence rationale annotation is used to guide the learning of token explainer $g_t(\cdot)$ through the intermediate sentence explainer $g_s(\cdot)$, leading to that token rationales come from true evidence, rather than from noise evidence. To this end, we adopt the sentence rationale label to guide the training of sentence explainer $g_s(\cdot)$ by formulating it as a multi-label classification problem using cross entropy (CE) loss (Paranjape et al., 2020),

$$\mathcal{L}_{SS} = \text{CE}(\boldsymbol{m}, \boldsymbol{E}), \tag{7}$$

where $\boldsymbol{E} = \{E_i \in \{0, 1\}\}|_{i=0}^{n}$ denotes whether the sentence is annotated rationale by humans.

(II) Salience-Token: As without direct guidance, to avoid the hard convergence of the token explainer $g_t(\cdot)$ during training, we construct the salience score $\boldsymbol{S} = \{\boldsymbol{s}_i\}|_{i=0}^{n}$ as soft reference for each token within each piece of evidence via the Layered Integrated Gradient (Kokhlikyan et al., 2020), where $\boldsymbol{s}_i = \{s_{i,j} \in [-1, 1]\}|_{j=0}^{|\boldsymbol{x}_i|}$. Then the KL divergence is employed to regularize the token rationale extraction,

$$\mathcal{L}_{ST} = \sum_{i=0}^{n} \text{KL}(P(\boldsymbol{z}_i)||\hat{\boldsymbol{s}}_i), \tag{8}$$

where $P(\boldsymbol{z}_i) = \text{softmax}_j(pt_{i,j})$ denotes the importance score of tokens over the i-th evidence, $\hat{\boldsymbol{s}}_i = \text{softmax}_j(s_{i,j})$.

In addition, to regularize the *compactness* of token rationale as (Jiang et al., 2021), we minimize the number of non-zeros predicted by the token explainer $g_t(\cdot)$ via minimizing the $\mathcal{L}_0$ norm with expectation (De Cao et al., 2020),

$$\mathcal{L}_0 = \sum_{i}^{n} \sum_{j}^{|\boldsymbol{x}_i|} pt_{i,j}. \tag{9}$$

3.4 Optimization

The optimization objective is minimizing the following loss function $\mathcal{L}$,

$$\mathcal{L} = \lambda_1 \mathcal{L}_F + \lambda_2 \mathcal{L}_C + \lambda_3 \mathcal{L}_{SS} + \lambda_4 \mathcal{L}_{ST} + \lambda_5 \mathcal{L}_0, \tag{10}$$

where λ_{1-5} are hyperparameters.

During training, we freeze the parameters of the pretrained veracity prediction module $f(\cdot)$ and only optimize the explainer parameters (i.e., $g_t(\cdot)$ and $g_s(\cdot)$) by minimizing $\mathcal{L}$. In the inference stage, the value of $z_{i,j}$ and m_i are determined by $\mathbb{1}(pt_{ij} > \alpha)$ and $\mathbb{1}(ps_i > \alpha)$, respectively, where α is the threshold of rationales, $\mathbb{1}(\cdot)$ is the indicator function.

4 Experiments

Datasets We perform experiments on three multi-hop fact verification datasets, including HoVer (Jiang et al., 2020), LIAR-PLUS (Alhindi et al., 2018), and PolitiHop (Ostrowski et al., 2021). For HoVer, following (Khattab et al., 2021), the dataset is constructed with retrieved evidence, where each claim is associated with 5 pieces of evidence. For LIAR-PLUS and PolitiHop, we use the datasets provided in Ostrowski et al. (2021) and restrict each claim associated with 10 and 5 pieces of evidence, respectively. All the datasets require multi-hop reasoning and consist of annotated sentence-level rationale and noise evidence.

Baselines Since no other works aimed at multi-granular rationale extraction, we compare CURE with twelve single-granular rationale extraction methods, including eight intrinsic-based methods (i.e., Pipeline in ERASER (DeYoung et al., 2020), Information Bottleneck (IB) (Paranjape et al., 2020), Two-Sentence Selecting (TSS) (Glockner et al., 2020), Learning from rationales (LR) (Carton et al., 2022) for sentence rationale extraction. Lei et al. (2016), DeClarE (Popat et al., 2018), FRESH (Jain et al., 2020), VMASK (Chen and Ji, 2020) for token rationale extraction.) and four post hoc methods (i.e., LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), Layer Integrated Gradient (L-INTGRAD) (Mudrakarta et al., 2018), DIFFMASK (De Cao et al., 2020) for token rationale extraction).

Metrics Inspired by DeYoung et al. (2020), we adopt the macro F1 and accuracy for verification prediction evaluation, and macro F1, Precision and Recall to measure the sentence-level agreement with human-annotated rationales. We also report fidelity defined in Eq.5 as a metric of faithfulness for post hoc methods. Moreover, we propose a new metric named **Noiselessness** based on the *Definition2* to measure the degree to which the token rationales contain noise[2]. It can be calculated by Token Rationale Overlap rate, which measures the overlap between token rationale with sentence Rationale (**TRO-R**) or Non-rationale (**TRO-N**),

$$\begin{aligned}
\text{TRO-R} &:= \frac{1}{|X_s|} \sum_{\boldsymbol{x_i} \in X_s} \frac{|\boldsymbol{x_i} \cap \boldsymbol{r_i}|}{|\boldsymbol{x_i}|}, \\
\text{TRO-N} &:= \frac{1}{|X_{ns}|} \sum_{\boldsymbol{x_i}' \in X_{ns}} \frac{|\boldsymbol{x_i}' \cap \boldsymbol{r_i}|}{|\boldsymbol{x_i}'|}, \\
\text{Noiselessness} &:= 1 - \frac{\text{TRO-N}}{\text{TRO-R}},
\end{aligned} \tag{11}$$

where $X_{ns} = X \setminus X_s$ denotes the complement subset of X_s. Higher Noiselessness with higher accuracy means that token rationales contain more true tokens from true evidence, rather than noise evidence, i.e., less noise and high purity.

Implementation Details Our Veracity Prediction model adopts the pretrained RoBERTa (Liu et al., 2019) base model to initialize the Transformer components and three hop steps are used (i.e., $L = 3$). The maximum number of input tokens to RoBERTa is 130 and the dimension of class capsule d_c is 10. The pretrained model has 80.03%, 83.14%, and 71.63% on label accuracy of claim verification on HoVer, LIAR-PLUS, and PolitiHop, respectively.

5 Results and Discussion

5.1 Quantitative Analysis

Main results Tab.1 presents the results from our CURE against the baselines for claim verification and rationale extraction. We report our main evaluation of the multi-granular rationale extraction on CURE*. Moreover, since the baselines cannot

[2]This metric should be considered together with the evaluation of claim verification to avoid spurious high noiselessness.

Dataset	Model	Rationale		Claim	
		Noiselessness↑	Fidelity↓	Acc.	F1
LIAR-PLUS	Lei et al. (2016)	0.2619	-/-	0.5681	0.5442
	DeClarE	-0.0030	-/-	0.4773	0.2154
	FRESH	0.2921	-/-	0.4345	0.4137
	VMASK	-0.0203	-/-	**0.8262**	**0.8146**
	LIME	0.4819	0.8422	0.3061	0.1562
	SHAP	0.0111	1.6401	0.7639	0.7531
	L-INTGRAD	0.0210	0.5889	0.7172	0.6984
	DIFFMASK	0.0131	4.2244	0.5850	0.4803
	CURE*	**0.6301**	**0.2675**	0.8210	0.8078
	CURE	0.5202	0.2675	0.8210	0.8078
	CURE *-C*	0.2641	0.2642	0.8132	0.8028
	CURE *-SS*	0.3968	0.3101	0.7704	0.7502
	CURE *-ST*	0.3754	0.2541	0.8171	0.8069
HoVer	Lei et al. (2016)	0.0667	-/-	0.5015	0.3410
	DeClarE	0.2063	-/-	0.5083	0.5076
	FRESH	-0.7372	-/-	0.6028	0.6014
	VMASK	0.3720	-/-	0.7438	0.7369
	LIME	0.7077	0.8356	0.5000	0.3333
	SHAP	0.0870	2.6125	0.5983	0.5818
	L-INTGRAD	0.1912	0.7058	0.5003	0.5386
	DIFFMASK	0.6474	1.1632	0.7153	0.7130
	CURE*	**0.9522**	**0.2287**	**0.7698**	**0.7689**
	CURE	0.7456	0.2287	0.7698	0.7689
	CURE *-C*	0.7381	0.2405	0.7585	0.7561
	CURE *-SS*	0.4978	0.3469	0.7298	0.7297
	CURE *-ST*	0.7476	0.2330	0.7683	0.7672
PolitiHop	Lei et al. (2016)	0.1489	-/-	0.5674	0.3691
	DeClarE	0.1053	-/-	0.6950	0.2734
	FRESH	0.1505	-/-	0.6170	0.4435
	VMASK	0.0075	-/-	**0.7234**	**0.5580**
	LIME	0.0057	0.8041	0.6950	0.2734
	SHAP	-0.0146	2.2659	0.5957	0.4071
	L-INTGRAD	0.0022	0.6580	0.6950	0.2734
	DIFFMASK	0.0134	2.4533	0.6738	0.4471
	CURE*	**0.8704**	**0.3204**	0.6950	0.3236
	CURE	0.4092	0.3204	0.6950	0.3236
	CURE *-C*	0.3735	0.2984	0.6525	0.3553
	CURE *-SS*	0.1479	0.2563	0.6950	0.4214
	CURE *-ST*	0.3877	0.3372	0.6809	0.2951

Table 1: Evaluation results of multi-granular rationale across three datasets. **CURE*** denotes the results using predicted token rationale and predicted sentence rationale, **CURE** denotes the results using predicted token rationale and annotated sentence rationale. ↑ means the larger value is better. $-C$, $-SS$ and $-ST$ denote the constraint removal of *Consistency*, *Salience-Sentence* and *Salience-Token*, respectively. Our main results are marked in bold and the best results are underlined.

extract the two granular rationales simultaneously, for a fair comparison, we also report the evaluation using the sentence rationale annotated by humans instead of the predicted sentence rationale to compute the Noiselessness. We can observe that: (I) CURE is capable of extracting **noiseless** rationales with the highest Noiselessness score, which indicates the importance of the differential between true evidence and noise evidence for the token rationale extraction. This is significantly reflected in the CURE*. However, in contrast, all baselines fail to induce noiseless rationales, leaving a significant gap with our CURE. (II) CURE is quite **faithful** with the lowest fidelity value across all three datasets, surpassing all other baselines. This result is in accordance with Jiang et al. (2021) that the Euclidean distance between the logits constrains the explainer to provide more faithful explanations. (III) On claim verification, our CURE outperforms the post hoc methods, while slightly lower compared with intrinsic methods. We conjecture that the information leakage caused by soft selection may improve the performance of these models.

Beyond relative performance against baselines, we conduct control experiments in the ablation study to explore the effectiveness of **property**. With the removal of different properties individually, we observe the reduced performance in the extracted rationales, both in fidelity and noiselessness. The most significant property is Salience-Sentence,

Dataset	Model	Sentence Rationale			Claim	
		F1	Precision	Recall	Acc.	F1
LIAR-PLUS	**Pipeline**	0.6677	0.7450	0.6564	0.5811	0.5393
	IB	0.3777	0.3927	0.3967	0.6252	0.6048
	TSS	0.4324	0.6349	0.3469	0.6239	0.6172
	LR	0.6242	0.6776	0.6381	0.7652	0.7519
	CURE	**0.6789**	**0.8072**	**0.6329**	**0.8210**	**0.8078**
HoVer	**0.9427**	0.9028	0.9900	**Pipeline**	0.6255	0.6244
	IB	0.6236	0.7018	0.5783	0.5678	0.5674
	TSS	0.6883	0.9026	0.5755	0.5368	0.5111
	LR	0.9419	0.9029	**0.9988**	0.5110	0.4050
	CURE	0.9376	**0.9045**	0.9877	**0.7698**	**0.7689**
PolitiHop	**Pipeline**	0.6390	0.5986	0.8234	0.6596	0.4173
	IB	0.4180	0.5106	0.3902	0.6879	**0.5489**
	TSS	0.4272	0.5177	0.4044	0.6525	0.4334
	LR	0.5699	0.5674	0.6657	**0.7021**	0.4712
	CURE	**0.6947**	**0.6584**	**0.8403**	0.6950	0.3459

Table 2: Evaluation of claim verification and sentence rationale extraction across three datasets. The best results are marked in bold.

Model	Spearman	F1	Precision	Recall
LIME	0.1695	0.5422	0.6564	0.5459
SHAP	0.0305	0.3636	0.5138	0.5170
L-INTGRAD	0.0776	0.5108	0.5314	0.5479
VMASK	0.1177	0.5247	0.5473	0.5732
CURE	**0.4293**	**0.6747**	**0.6739**	**0.7650**

Table 3: Evaluation of token rationale extraction on the HoVer dataset based on our re-annotation. The best results are marked in bold.

this can be due to that explainer is susceptible to over-fitting and yields noise token rationales from noise sentences when lacking the ground-truth information of sentence rationales. The second key property is Consistency, there are varying decreases in both fidelity and noiselessness throughout the three datasets, particularly for LIAR-PLUS, which requires more complex rationales for reasoning over multiple pieces of evidence than the other two datasets. We reasonably presume the synergy of the two granular explainers by constraining the extraction of *right token* from *right sentence* (Gupta et al., 2022). Moreover, we note a minor decrease in claim verification when removing the Salience-Token, showing that the retained task-relevant tokens directed by the salience score can help boost the performance of veracity prediction.

Plausibility As shown in Tab.2 and Tab.3, we further conduct the experiments to explore how well the extracted rationales agree with human annotation (Jacovi and Goldberg, 2020) compared to classical single-granular rationale methods.

For **sentence rationale**, surprisingly, we find that our CURE still outperforms the most baselines on claim verification and rationale extraction. We reasonably posit that the high quality *right token* is useful for extracting *right sentence* rationale in turn. To further validate the quality of **token rationale** extraction, we ask 3 annotators with NLP backgrounds to re-annotate 150 samples from the development set of the HoVer dataset to obtain the token rationale label. Our annotators achieve 0.6807 on Krippendorff's α (Krippendorff, 2011) and retain 20% tokens annotated as rationales for each sample. We measure the agreement between the predicted token rationales and human annotated rationales with the *Spearman's correlation*, *macro F1*, *Precision*, and *Recall*. As shown in Tab.3, our CURE is far more promising and outperforms the baselines with a huge gap on all evaluation metrics. It clearly demonstrates the effectiveness of the denoising rationalization framework we proposed for explaining multi-hop fact verification.

5.2 Manual Evaluation

Inspired by Zhou et al. (2020) and Yan et al. (2022), we provide a manual evaluation of the token rationales (contained in the sentence rationale rather than all the evidence) extracted by CURE, compared to DIFFMASK (De Cao et al., 2020) and VMASK (Chen and Ji, 2020). We randomly select 50 samples and ask three annotators with NLP backgrounds to score these rationales in a likert scale of 1 to 5 according to three different criteria: (I) **Correctness**, which measures what extent users can approach ground-truth label given the predicted token rationales; (II) **Faithfulness**, which measures what extent users can approach the model predicted label given the predicted token rationales;

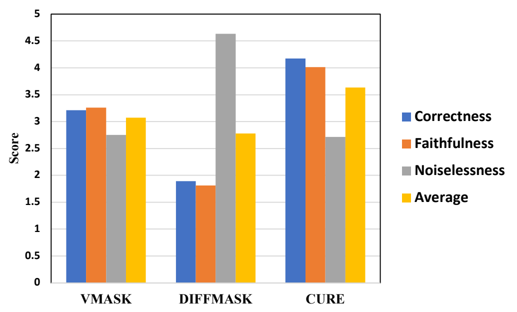

Figure 4: Human evaluation results in a likert scale of 1 to 5, where 1 means strongly disagree and 5 means strongly agree. *Average* denotes the average score of three criteria. The inner-rater agreement measured by Krippendorff's α is 0.88.

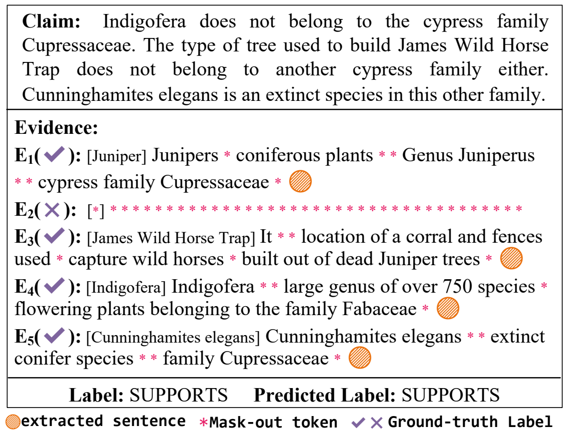

Figure 5: Rationales extracted by our method on the HoVer dataset. All the tokens except * denote the token rationales predicted by our CURE.

(III) **Noiselessness**, which measures what extent the predicted token rationales do not contain noise and irrelevant words.

The human evaluation results are shown in Fig.4. We can observe that CURE achieves the best results on correctness and faithfulness. Although DIFFMASK performs particularly well on noiselessness, the correctness and faithfulness of the generated rationales are far worse than those of the other two models, indicating the low quality of its rationales. In fact, DIFFMASK excels at masking almost all tokens due to the only constraint of L_0 loss. Considering the mutual constraints between noiselessness and the other two criteria, we calculate the average scores of three criteria for each method. CURE still outperforms on average score, which demonstrates the high quality of the noiseless token rationales generated by our method.

5.3 Rationale Examples

Fig. 5 presents an intuitive example with rationales generated by our CURE from the HoVer dataset. We can observe that our CURE correctly predicts the sentence rationales while entirely removing the noise sentence $E2$. Meanwhile, The corresponding retained token rationales contain information that is not only important for veracity prediction, but also appears the true token rationales by ignoring the noise tokens, demonstrating their faithfulness and noiselessness. It is worth noting that our CURE is prone to retaining the *title* of the document as the key cue for linking multiple pieces of evidence.

6 Related Work

A growing interest in interpretability has led to a flurry of approaches in trying to reveal the reasoning behavior behind the multi-hop fact verification task. A well-studied way is to use the attention weights as attribution score to indicate the importance of a token or sentence, such as self-attention (Popat et al., 2018) or co-attention (Shu et al., 2019; Yang et al., 2019; Wu et al., 2020, 2021). While this method is incapable of guaranteeing the inattention of low-score features, drawing criticism recently (Wiegreffe and Pinter, 2019; Meister et al., 2021). Another line of research focuses on perturbation-based methods. These methods explore a built-in explainer to generate the rationale by masking the unimportant language features (Atanasova et al., 2020; Paranjape et al., 2020; Glockner et al., 2020; Kotonya and Toni, 2020b; Zhang et al., 2021; Fajcik et al., 2022). This way generally employs the *extract-then-predict* paradigm, while Yu et al. (2021) reveals an issue of model interlocking in such a cooperative rationalization paradigm.

Recently, a few studies have explored the post hoc paradigm for explanation extraction by detaching the explainer and the task model. With the parameters of the task model frozen, they focus on the external explainer to retain the key cue in input as the rationales to indicate features the task model relies on (De Cao et al., 2020; Si et al., 2023; Atanasova et al., 2022; Ge et al., 2022). Our work falls under the scope of the post hoc paradigm, different from the prior works that only consider the single-granular rationale, we for the first time propose a novel paradigm to yield indicative token rationales by regularizing the multi-granular rationale extraction.

7 Conclusion

We propose a novel multi-granular rationale extraction framework for denoising rationalization in multi-hop fact verification. The parallel explainers are collaboratively modeled by constraining with three diagnostic properties. A new noiselessness metric is introduced to measure the purity of the rationales. The results on three multi-hop fact verification datasets illustrate the effectiveness of our method. In the future, we will explore how to generate counterfactual explanations.

Limitations

A limitation of our work is that we employ the supervised paradigm because of the difficulty in satisfying our expectations about the rationales. We need the labels of sentence-level rationales as guidance to obtain better classification performance and high-quality rationales, which may be difficult to extend our method into the scenarios with few annotations (i.e., semi-supervised or unsupervised). In addition, the L_0 loss regularization overemphasizes the sparsity, which can damage the performance on claim verification and make the model sensitive to hyperparameters.

Acknowledgement

The authors would like to thank the anonymous reviewers for their insightful comments. This work is funded by the National Natural Science Foundation of China (Grant No.62176053, No.62402258, No.62376130), Shandong Provincial Natural Science Foundation (Grant No.ZR2024QF099), Program of New Twenty Policies for Universities of Jinan (Grant No.202333008), the Pilot Project for Integrated Innovation of Science, Education, and Industry of Qilu University of Technology (Shandong Academy of Sciences)(2024ZDZX08), and supported by the Big Data Computing Center of Southeast University.

References

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2017. Deep variational information bottleneck. In *International Conference on Learning Representations*.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2022. Diagnostics-guided explanation generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10445–10453.

Samuel Carton, Surya Kanoria, and Chenhao Tan. 2022. What to learn, and how: Toward effective learning from rationales. In *Proceedings of Findings of the Association for Computational Linguistics: ACL*, pages 1075–1088.

Hanjie Chen and Yangfeng Ji. 2020. Learning variational word masks to improve the interpretability of neural text classifiers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 4236–4251.

Nicola De Cao, Michael Sejr Schlichtkrull, Wilker Aziz, and Ivan Titov. 2020. How do decisions emerge across layers in neural models? interpretation with differentiable masking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3243–3255.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458.

Martin Fajcik, Petr Motlicek, and Pavel Smrz. 2022. Claim-dissector: An interpretable fact-checking system with joint re-ranking and veracity prediction. *arXiv preprint*, arXiv:2207.14116.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728.

Ling Ge, ChunMing Hu, Guanghui Ma, Junshuang Wu, Junfan Chen, JiHong Liu, Hong Zhang, Wenyi Qin, and Richong Zhang. 2022. E-VarM: Enhanced variational word masks to improve the interpretability of text classification models. In *Proceedings of the International Conference on Computational Linguistics*, pages 1036–1050.

Max Glockner, Ivan Habernal, and Iryna Gurevych. 2020. Why do you think that? exploring faithful sentence-level rationales without supervision. In *Proceedings of Findings of the Association for Computational Linguistics: EMNLP*, pages 1080–1095.

Vivek Gupta, Shuo Zhang, Alakananda Vempala, Yujie He, Temma Choji, and Vivek Srikumar. 2022. Right

for the right reason: Evidence extraction for trustworthy tabular reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 3268–3283.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Proceedings of Findings of the Association for Computational Linguistics: EMNLP*, pages 3441–3460.

Zhongtao Jiang, Yuanzhe Zhang, Zhao Yang, Jun Zhao, and Kang Liu. 2021. Alignment rationale for natural language inference. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 5372–5387.

Omar Khattab, Christopher Potts, and Matei Zaharia. 2021. Baleen: Robust multi-hop reasoning at scale via condensed retrieval. In *Advances in Neural Information Processing Systems*, volume 34, pages 27670–27682.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch. *CoRR*, abs/2009.07896.

Neema Kotonya and Francesca Toni. 2020a. Explainable automated fact-checking: A survey. In *Proceedings of the International Conference on Computational Linguistics*, pages 5430–5443.

Neema Kotonya and Francesca Toni. 2020b. Explainable automated fact-checking for public health claims. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 7740–7754.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 107–117.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220.

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.

Christos Louizos, Max Welling, and Diederik P. Kingma. 2018. Learning sparse neural networks through L_0 regularization. In *International Conference on Learning Representations*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2022. Towards faithful model explanation in nlp: A survey. *arXiv preprint*, arXiv:2209.11326.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*.

Clara Meister, Stefan Lazov, Isabelle Augenstein, and Ryan Cotterell. 2021. Is sparse attention more interpretable? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 122–129.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1896–1906.

Wojciech Ostrowski, Arnav Arora, Pepa Atanasova, and Isabelle Augenstein. 2021. Multi-hop fact checking of political claims. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3892–3898.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An information bottleneck approach for controlling conciseness in rationale extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1938–1952.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 22–32.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1135–1144.

Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. 2017. Dynamic routing between capsules. In *Proceedings of the International Conference on Neural Information Processing Systems*, page 3859–3869.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. Defend: Explainable fake news detection. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 395–405.

Jiasheng Si, Deyu Zhou, Tongzhe Li, Xingyu Shi, and Yulan He. 2021. Topic-aware evidence reasoning and stance-aware aggregation for fact verification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 1612–1622.

Jiasheng Si, Yingjie Zhu, and Deyu Zhou. 2023. Exploring faithful rationale for multi-hop fact verification via salience-aware graph learning. In *Proceedings of Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI*, pages 13573–13581.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning*, volume 70, pages 3319–3328.

Yixuan Tang, Hwee Tou Ng, and Anthony Tung. 2021. Do multi-hop question answering systems know how to answer the single-hop sub-questions? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3244–3249, Online. Association for Computational Linguistics.

Tijmen Tieleman, Geoffrey Hinton, et al. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 11–20.

Lianwei Wu, Yuan Rao, Yuqian Lan, Ling Sun, and Zhaoyin Qi. 2021. Unified dual-view cognitive model for interpretable claim verification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 59–68.

Lianwei Wu, Yuan Rao, Yongqiang Zhao, Hao Liang, and Ambreen Nazir. 2020. DTCA: Decision tree-based co-attention networks for explainable claim verification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1024–1035.

Hanqi Yan, Lin Gui, and Yulan He. 2022. Hierarchical interpretation of neural text classification. *arXiv preprint arXiv:2202.09792*.

Fan Yang, Shiva K. Pentyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D. Ragan, Shuiwang Ji, and Xia (Ben) Hu. 2019. Xfake: Explainable fake news detector with visualizations. In *The World Wide Web Conference*, page 3600–3604.

Mo Yu, Yang Zhang, Shiyu Chang, and Tommi Jaakkola. 2021. Understanding interlocking dynamics of cooperative rationalization. In *Advances in Neural Information Processing Systems*, volume 34.

Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. 2019. Lookahead optimizer: k steps forward, 1 step back. In *Advances in Neural Information Processing Systems*, volume 32.

Zijian Zhang, Koustav Rudra, and Avishek Anand. 2021. Explain and predict, and then predict again. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, page 418–426.

Wangchunshu Zhou, Jinyi Hu, Hanlin Zhang, Xiaodan Liang, Maosong Sun, Chenyan Xiong, and Jian Tang. 2020. Towards interpretable natural language understanding with explanations as latent variables. In *Advances in Neural Information Processing Systems*, volume 33, pages 6803–6814.

A Generalized multi-granular Faithfulness and Noiselessness

Multi-granular faithfulness and noiselessness can be easily applied to explain a model with sequential input.

For the given trained model $f(\cdot)$ and input $X = [\boldsymbol{c}; \boldsymbol{e_1}; \boldsymbol{e_2}; \ldots; \boldsymbol{e_n}]$ with its corresponding rationales $R_s = \{\boldsymbol{s_i} \mid \boldsymbol{s_i} \in \{\boldsymbol{e_1}, \boldsymbol{e_2}, \ldots, \boldsymbol{e_n}\}\}$ and $\boldsymbol{r} = \{\boldsymbol{r}_i \subset \boldsymbol{e}_i\}|_{i=0}^{n}$, where $\boldsymbol{r}_i = \{\boldsymbol{t}_{i,j} | \boldsymbol{t}_{i,j} \in \boldsymbol{e}_i\}$, we define the generalized multi-granular *faithfulness* and *noiselessness* of R_s and $\boldsymbol{r}$ as follows.

Definition 3. *(Faithfulness)* R_s *and* $\boldsymbol{r}$ *are multi-granular faithful to their corresponding prediction* Y *if and only if* Y *rely entirely on* $X_R = [\boldsymbol{c}, \boldsymbol{e}_1 \cap R_s \cap \boldsymbol{r}_1, \ldots, \boldsymbol{e}_n \cap R_s \cap \boldsymbol{r}_m]$.

Definition 4. *(Noiselessness) R_s and $\boldsymbol{r}$ are multi-granular noiseless to their corresponding prediction Y if and only if satisfying*

$$\sum_{\boldsymbol{s}_i \in R_s} |\boldsymbol{s}_i \cap \boldsymbol{r}_i| \leq \epsilon, \sum_{\boldsymbol{s}'_i \in X \setminus R_s} |\boldsymbol{s}'_i \cap \boldsymbol{r}_i| \to 0, \quad (12)$$

where ϵ is the maximum expected sparsity of token-level rationales, $X \setminus R_s$ denotes the complementary subset of R_s.

B Hard Concrete Distribution

The Hard Concrete distribution (Louizos et al., 2018) assigning probability densities on the close unit interval $[0, 1]$ by using *stretch* and *rectify* the binary Concrete distribution (Maddison et al., 2017). We describe this process here briefly.

Assume we have a binary concrete random variable s distributed in the interval $(0, 1)$ with its parameters $\phi = (\log \alpha, \beta)$, where $\log \alpha$ is the location and β is the temperature and probability density function (pdf) $q_s(s|\phi)$ and cumulative density function (CDF) $Q_s(s|\phi)$,

$$q_s(s|\phi) = \frac{\beta \alpha s^{-\beta-1}(1-s)^{-\beta-1}}{(\alpha s^{-\beta} + (1-s)^{-\beta})^2} \quad (13)$$

$$Q_s(s|\phi) = \sigma((\log s - \log(1-s))\beta - \log \alpha. \quad (14)$$

where $\sigma(\cdot)$ denotes Sigmoid$(\cdot)$.

We can stretch the above distribution to the interval (γ, ζ), with $\gamma < 0$ and $\zeta > 1$ and obtain

$$\begin{aligned} \bar{s} &= s(\zeta - \gamma) + \gamma, \\ s &= \sigma((\log u - \log(1-u) + \log \alpha)/\beta), \\ u &\sim U(0, 1), \end{aligned} \quad (15)$$

with the corresponding pdf and CDF

$$q_{\bar{s}}(\bar{s}|\phi) = \frac{1}{|\zeta| - \gamma} q_s(\frac{\bar{s} - \gamma}{\zeta - \gamma}|\phi), \quad (16)$$

$$Q_{\bar{s}}(\bar{s}|\phi) = Q_s(\frac{\bar{s} - \gamma}{\zeta - \gamma}|\phi). \quad (17)$$

Then, by further rectifying $\bar{s}$ with the hard-sigmoid,

$$z = \min(1, \max(0, \bar{s})), \quad (18)$$

we can obtain a distribution over z:

$$\begin{aligned} q(z|\phi) = Q_{\bar{s}}(0|\phi)\delta(z) + (1 - Q_{\bar{s}}(1|\phi))\delta(z-1) \\ + (Q_{\bar{s}}(1|\phi) - Q_{\bar{s}}(0|\phi))q_{\bar{s}}(z|\bar{s} \in (0,1), \phi). \end{aligned} \quad (19)$$

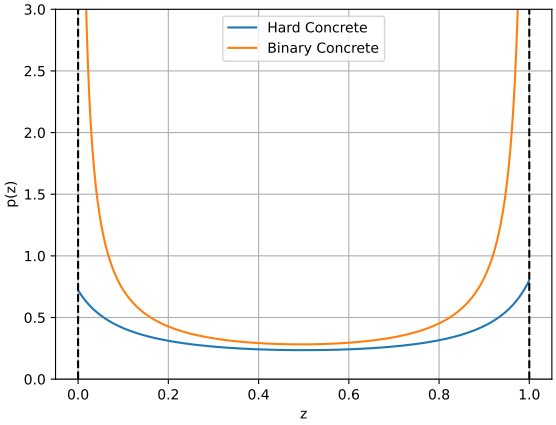

Figure 6: The binary concrete distribution with location $\log \alpha = 0$ and temperature $\beta = 0.3$. The hard concrete distribution is obtained by stretching the binary concrete distribution to $(\gamma = -0.1, \zeta = 1.1)$ and then applying a hard-sigmoid.

as Fig.6 shows, which is composed of a delta peak at zero with probability $Q_{\bar{s}}(0|\phi)$, a delta peak at one with probability $1 - Q_{\bar{s}}(1|\phi)$, and a truncated version of $q_{\bar{s}}(\bar{s}|\phi)$ in the $(0, 1)$ range.

According to Eq16, we can derive the expectation of non-zero for z,

$$p = \int_0^1 z q(z|\phi)\, \mathrm{d}z + q(1|\phi), \quad (20)$$

which is treated as the importance score of each token or sentence in the input.

C Capsule Network with Intervention

We take the sentence representation $\tilde{\boldsymbol{h}}^L_{i,0}|^n_{i=0}$ as evidence capsules $\boldsymbol{pc}_i|^n_{i=0} \in \mathbb{R}^d$ and the label as the class capsule $\boldsymbol{cc}_j|^m_{j=1} \in \mathbb{R}^{d_c}$, where m denotes the number of classes. The capsule network models the relationship between evidence capsules and class capsules by computing the routing weights using a dynamic routing mechanism.

Specifically, let $\boldsymbol{pc}_{j|i}$ be the predicted vector for the pair of evidence capsule $\boldsymbol{pc}_i$ and the class capsule $\boldsymbol{cc}_j$,

$$\boldsymbol{pc}_{j|i} = \boldsymbol{W}_{ji} \boldsymbol{pc}_i, \quad (21)$$

where $\boldsymbol{W}_{ji} \in \mathbb{R}^{d_c \times d}$ denotes the transformation matrix. Then, all the evidence capsules are aggregated by a weighted summation over all corresponding predicted vectors to generate the presentation of each class capsule:

$$\boldsymbol{cc}_j = Squash(\sum_i^n \beta_{ji} \boldsymbol{pc}_{j|i}), \quad (22)$$

Dataset		Num.	C.	Avg.L	Avg.SR
HoVer	**Train**	18171	2	66.28	3.04
	Dev	4000	2	53.66	3.37
LIAR-PLUS	**Train**	6341	3	23.49	3.89
	Dev	771	3	23.34	3.93
PolitiHop	**Train**	592	3	24.92	2.53
	Dev	141	3	25.41	2.55

Table 4: Statistics of datasets. *Num.* and *C.* are the number of claims and classes of each dataset, respectively. *Avg.L* denotes the average length of evidence, and *Avg.SR* denotes the average number of sentence rationales.

where $Squash(\cdot)$ is a non-linear squashing function which limits the length of $\boldsymbol{cc}_j$ to $[0, 1]$ and β_{ji} denotes the coupling coefficient and is calculated by iterative dynamic routing algorithm on original logits b_{ji}, which is summarized in Algorithm 1. Finally, the claim can be classified by choosing the class capsule with the largest ρ_j via optimizing the capsule loss.

To further eliminate the effect of noise sentences, we use sentence mask vector $\boldsymbol{m}$ to intervene the coupling coefficients β_{ij} between the evidence capsule $\boldsymbol{pc_i}$ and class capsule $\boldsymbol{cc_j}$ in the dynamic routing, as described in line 5 of Algorithm 1.

Algorithm 1 Dynamic Routing Algorithm

1: **Procedure:** Routing($\boldsymbol{pc}_{j|i}, \hat{p}_{ji} = |\boldsymbol{pc}_i|$)
2: Initialize the coupling coefficient $b_{ij} \rightarrow 0$
3: **for** iteration **do**
4: For all capsules: $\beta_{ji} \leftarrow \hat{p}_{ji} \cdot \text{leaky\_softmax}(b_{ji})$
5: For all capsules: $\beta_{ji} \leftarrow \beta_{ji} \cdot m_i$
6: For class capsules: $s_j \leftarrow \sum_i \beta_{ji} \cdot \boldsymbol{pc}_{j|i}$
7: For class capsules: $\boldsymbol{cc_j} \leftarrow Squash(s_j)$
8: For all capsules: $b_{ji} \leftarrow b_{ji} + \boldsymbol{pc}_{j|i} \cdot \boldsymbol{cc}_j$
9: **end for**
10: **return** $\boldsymbol{cc} \in \mathbb{R}^{m \times d_c}, \rho_j = |\boldsymbol{cc}_j|$

D Experimental Setup

D.1 Datasets

All the datasets we used require complex reasoning over multiple pieces of evidence and have annotated sentence rationales by humans. Note that evidence of the datasets has a large portion of noise evidence, leading to confusion for the model in the reasoning process. For HoVer, given a claim with 5 corresponding evidence sentences, a fact verification model needs to predict the veracity of the claim $\in \{SUPPORTS, REFUTES\}$. For LIAR-PLUS and PolitiHop, given a claim with 10 and 5 corresponding evidence sentences, respectively, a fact verification model needs to predict the veracity of the claim $\in \{false, half-true, true\}$. Furthermore, as Section 5.1 described, we manually annotated the token-level rationales for 150 samples on the development set of HoVer. For each dataset, we used a script to check for and remove offensive content and identifiers. The statistics of the datasets are shown in Tab.4.

D.2 Baselines

- **Pipeline** (DeYoung et al., 2020): An *intrinsic* pipeline *sentence-level* rationale extraction method which consists of an extractor and a classifier and verdicts the claim using one sentence-level rationale with the highest confidence score from the extractor.

- **Information Bottleneck (IB)** (Paranjape et al., 2020): An *intrinsic sentence-level* rationale extraction model which employ the variational information bottleneck (Alemi et al., 2017) to extract rationales for prediction and interpretability simultaneously. We set the threshold to 0.6, 0.4 and 0.4 for HoVer, LIAR-PLUS, and PolitiHop, respectively.

- **Two-Sentence Selecting (TSS)** (Glockner et al., 2020): An *intrinsic sentence-level* rationale extraction model which extracts rationales by using the loss logits of sentences. We chose the first two sentences as sentence-level rationales because of the expensive cost of computing.

- **Learning from rationales** (Carton et al., 2022): An *intrinsic sentence-level* rationale extraction model with several novel loss functions and learning strategies, which aims to improve the performance of both prediction and rationale extraction.

- **LIME** (Ribeiro et al., 2016): A *post hoc token-level* rationale extraction method which computes the attribution score for each input token by using a linear model to locally approximate the model's predictions on a set of perturbed instances around the base data point. We use the implementation provided by Captum (Kokhlikyan et al., 2020) and choose LASSO

Veracity Prediction	**HoVer**	**LIAR-PLUS**	**PolitiHop**
Layers	12	12	12
Hidden size	768	768	768
Heads in GAT	8	8	12
Dropout rate in GAT	0.0	0.0	0.6
Optimizer	AdamW	AdamW	AdamW
Learning rate	1e-5	1e-5	1e-6
Train epochs	20	30	100
Batchsize	1	1	1
Rationale Extraction	**HoVer**	**LIAR-PLUS**	**PolitiHop**
Optimizer	Lookahead RMSprop*	Lookahead RMSprop*	Lookahead RMSprop*
Learning rate g_s g_t	3e-4	3e-4	3e-4
Train epochs	10	20	100
λ_{1-5}	$\{1, 0.1, 1, 0.15, 0.2\}$	$\{1, 0.2, 1, 0.1, 0.1\}$	$\{1, 0.1, 1, 0.15, 0.45\}$

Table 5: Hyperparameters for training. Optimizer: *Tieleman et al. (2012); Zhang et al. (2019).

Dataset	**Model**	**Claim Verification**		**Sentence-level Rationale**		
		Acc.	**F1**	**F1**	**Precision**	**Recall**
LIAR-PLUS	**CURE**	**0.8210**	**0.8078**	**0.6789**	**0.8072**	**0.6329**
	CURE -*C*	0.8132	0.8028	0.6935	0.8001	0.6627
	CURE -*ST*	0.8171	0.8069	0.6766	0.8056	0.6326
HoVer	**CURE**	**0.7698**	**0.7689**	**0.9376**	**0.9045**	**0.9877**
	CURE -*C*	0.7585	0.7561	0.9383	0.9044	0.9892
	CURE -*ST*	0.7683	0.7672	0.9349	0.9066	0.9807
PolitiHop	**CURE**	**0.6950**	**0.3459**	**0.6947**	**0.6584**	**0.8403**
	CURE -*C*	0.6525	0.3553	0.7121	0.6565	0.8811
	CURE -*ST*	0.6809	0.2961	0.6954	0.6593	0.8397

Table 6: Ablation study results of sentence rationale extraction across three datasets. $-C$ and $-ST$ denote the constraint removal of *Consistency* and *Salience-Token*, respectively.

as the surrogate model. We consider all tokens with an attribution score greater than 0 as rationales.

- **Lei et al. (2016)**: An *intrinsic token-level* rationale extraction model which combines two components, generator and encoder, where the generator specifies a distribution over tokens as candidate rationales, and these are passed through the encoder for prediction. We implement generator training using the Gumbel Softmax instead of using REINFORCE for stable training [3].

- **SHAP** (Lundberg and Lee, 2017): A *post hoc token-level* rationale extraction method similar to LIME, but where the weights of each perturbed instance are computed based on Shapely values when training the linear model.

- **Layer Integrated Gradient (L-INTGRAD)** (Mudrakarta et al., 2018): A *post hoc token-level* rationale extraction method which is a variant of Integrated Gradients (Sundararajan et al., 2017) and assigns an importance score to layer inputs or outputs, depending on whether we attribute to the former or to the latter one. We use the implementation provided by Captum (Kokhlikyan et al., 2020) and consider all tokens with an attribution score greater than 0 as rationales.

- **DeClarE** (Popat et al., 2018): An *intrinsic token-level* rationale extraction model which extracts the rationales with the attention score. We choose the attention score with the threshold of 0.5.

- **FRESH** (Jain et al., 2020): A pipeline *intrinsic token-level* rationale extraction model which is a simpler variant of Lei et al. (2016),

[3] https://github.com/yala/text_nn

where rationales are induced by using arbitrary attribution scores (e.g., gradient from a trained model) heuristically. The words with the top 50%, 30%, and 40% attribution scores were selected as the rationales for HoVer, LIAR-PLUS, and PolitiHop, respectively, to ensure the comparable sparsity with our model results.

- **VMASK** (Chen and Ji, 2020): An *intrinsic token-level* rationale extraction model based on variational information bottleneck, which is similar to IB (Paranjape et al., 2020). We take tokens with an expectation greater than 0.5 as rationales.

- **DIFFMASK** (De Cao et al., 2020): A *post hoc token-level* rationale extraction model which learns to mask out a subset of the input tokens while maintaining a distribution overprediction as close to the original distribution as possible.

All experimental setups of the baselines are followed from the original papers.

D.3 Implementation Details

We report the hyperparameters for training veracity prediction models and multi-granular explainers in Tab.5. We run all the experiments using the Nvidia GeForce RTX 3090 (24 GB) GPU.

E Sentence Rationale Extraction Ablation Study

In the ablation study, we explore whether the constraints of **Consistency** and **Salience-Token** from the token rationale affect the performance of sentence rationale extraction. As shown in Tab.6, as expected, we observe a slight improvement in the performance of sentence rationale extraction and a slight decrease in claim verification when removing the consistency. On one hand, the claim verification will be affected by the noise tokens contained in the extracted sentence without the regularization of consistency from the token rationale. On the other hand, the consistency is a trade-off term between the token rationale extraction and sentence rationale extraction, thus the model will increase the performance of sentence rationale extraction by only focusing on the label information without the constraint from the consistency.

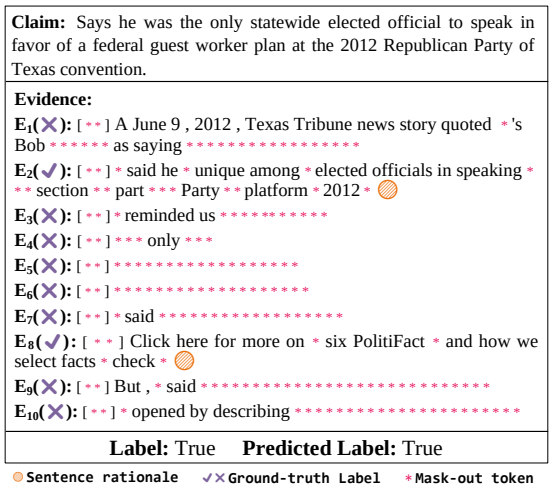

Figure 7: Rationales extracted by our method on LIAR-PLUS. Each piece of evidence consists of the author and an evidence sentence. All the retained tokens are the token rationales predicted by our method.

F Token Annotation Guidelines

We ask three annotators with NLP backgrounds to re-annotate the token rationale according to the following guidelines. The annotators are three graduate-level language technology researchers. We follow the local laws and offer 30 dollars per hour for each annotator.

Guidelines: Please determine whether each token in the evidence supports/refutes the claim according to the claim, the evidence, the classification label (SUPPORTS/REFUTES), and the sentence label of whether each evidence is a sentence-level rationale (1/0). Please read the following **detailed guidelines** and the corresponding **example** carefully:

- Please try to focus on the tokens that supports/refutes the claim, i.e., important tokens, preferably **requiring some inference** rather than simply using tokens that duplicate those in the claim as rationales.

- Please focus on the **useful evidence** (whose label is 1), i.e., the sentence-level rationales, rather than the useless evidence (whose label is 0).

- If you think that **part of the tokens** of a word is important, then you can label part of the tokens of the word as rationales.

585: Nicole Provis's had a partner in the 1992 Dow Classic–Doubles. Her partner has won more championship doubles titles than Vasek Pospisil.																			
label: SUPPORTS																			
</s>1992 Dow Classic – Doubles</s></s>Nicole Provis and Elizabeth Smylie were the defending champions but were defeated in the quarterfinals by Jo-Anne Faull and Julie Richardson .</s> **label: 1** </s>Jack Sock</s></s>A former junior US Open champion, Sock's singles success is highlighted by 7 ATP finals , including three titles .</s> **label: 0** </s>Elizabeth Smylie</s></s>During her career, she won four Grand Slam titles, one of them in women 's doubles and three in mixed doubles .</s> **label: 1** </s>Elizabeth Smylie</s></s>Elizabeth Smylie (née Sayers, born 11 April 1963), sometimes known as Liz Smylie, is a retired Australian professional tennis player .</s> **label: 1** </s>Vasek Pospisil</s></s>Along with partner Jack Sock, he won the 2014 Wimbledon Championships and the 2015 Indian Wells Masters men's doubles titles .</s> **label: 1**																			
</s>	1992	ĠDow	ĠClassic	ĠâĢ	ĠDou	bles	</s>	</s>	Nic	ole	ĠPro	vis	Ġand	ĠElizabeth	ĠSmy	lie	Ġwere	Ġthe	.....
0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	.....
</s>	Jack	ĠS	ock	</s>	</s>	A	Ġformer	Ġjunior	ĠUS	ĠOpen	Ġchampion	Ġ,	ĠS	ock	Ġ'	s	Ġsingles	Ġsuccess	.....
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.....
</s>	Elizabeth	ĠSmy	lie	</s>	</s>	During	Ġher	Ġcareer	Ġ,	Ġshe	Ġwon	Ġfour	ĠGrand	ĠSlam	Ġtitles	Ġ,	Ġone	Ġof	.....
0	1	1	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1	.....
</s>	Elizabeth	ĠSmy	lie	</s>	</s>	Elizabeth	ĠSmy	lie	Ġ(	Ġn	Ã©	ĠS	ayers	Ġ,	Ġborn	Ġ11	ĠApril	Ġ1963	.....
0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	.....
</s>	V	ase	k	ĠP	osp	is	il	</s>	</s>	Along	Ġwith	Ġpartner	ĠJack	ĠS	ock	Ġ,	Ġhe	Ġwon	.....
0	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	.....

Table 7: The example of token-level rationales annotation

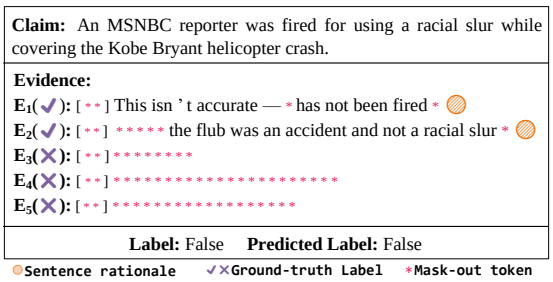

Figure 8: Rationales extracted by our method on PolitiHop. Each piece of evidence consists of the speaker and an evidence sentence. All the retained tokens are the token rationales predicted by our method.

- If you feel that **punctuation and special tokens </s>** are also important, please remember to label them as rationales.
- For samples with the classification label **REFUTES**, some parts of the claim may be correct and some may be incorrect. Please label **all the tokens** that can verify both the correct and incorrect parts as rationales.
- In some cases, you can find **errors in the ground-truth judgment** (i.e., the classification label or sentence label) or the evidence **does not contain enough information** to decide whether the claim should be supported or refuted. If you notice so, please skip this sample and mark it as *ERROR*.

That is all. Tab.7 is a specific example for annotation. Thank you for annotating!

G Rationale Examples

Fig.7 and Fig.8 present the intuitive examples on LIAR-PLUS and PolitiHop, respectively. The two datasets are PolitiFact-based datasets, which consist of the author profile in their evidence text. Unlike the title of the document, the profile of the author may trigger bias for the model. So our method does not consider the author or speaker to be rationales. It should be noted that each piece of evidence in LIAR-PLUS comes from a passage by the same author, so it is necessary to rely on other evidence (i.e. context) to determine whether each piece of evidence is correct. For example, in Fig.7, we need E_8 (probably a hyperlink to a PolitiFact article, an official fact-checking site) to prove the correctness of E_2, and further approach the veracity of the claim.