LOREN: Logic-Regularized Reasoning for Interpretable Fact Verification

Jiangjie Chen[1,2], Qiaoben Bao[1], Changzhi Sun[2], Xinbo Zhang[2], Jiaze Chen[2], Hao Zhou[2], Yanghua Xiao[1] and Lei Li[3]

[1]Fudan University [2]ByteDance AI Lab [3]University of California, Santa Barbara

Introduction

❖**The Fact Verification Task (FEVER)**

Verifying the veracity of a textual claim with evidence from trustworthy knowledge bases (e.g., Wikipedia).

- Misinformation detection on social media
- Factually accurate language generation

But... Why?

❖**Interpretable Fact Verification**

- Right answer for the right thinking

❖**Interpretability** "may be" the right thinking

- **Faithful**: able to explain the prediction
- **Accurate**: should be right per se
- **Debuggable**: able to find out where goes wrong

❖**The Research Problem**:

- How can we do it without supervision?

Motivation from Humans

Claim: c Donald Trump won the 2020 election.

❖We carefully examine each phrase in a claim one by one.

- Did Donald Trump win the election in **[2020]**?
- Did Donald Trump win the **[U.S.]** presidential election in 2020?

❖We aggregate the verification results of each phrase **following aggregation logic**, i.e. a claim is found

- ***Supported*** iff all phrases found supported;
- ***Refuted*** iff exists a phrase found refuted;
- ***NEI*** iff not refuted and exists a phrase found unverifiable.

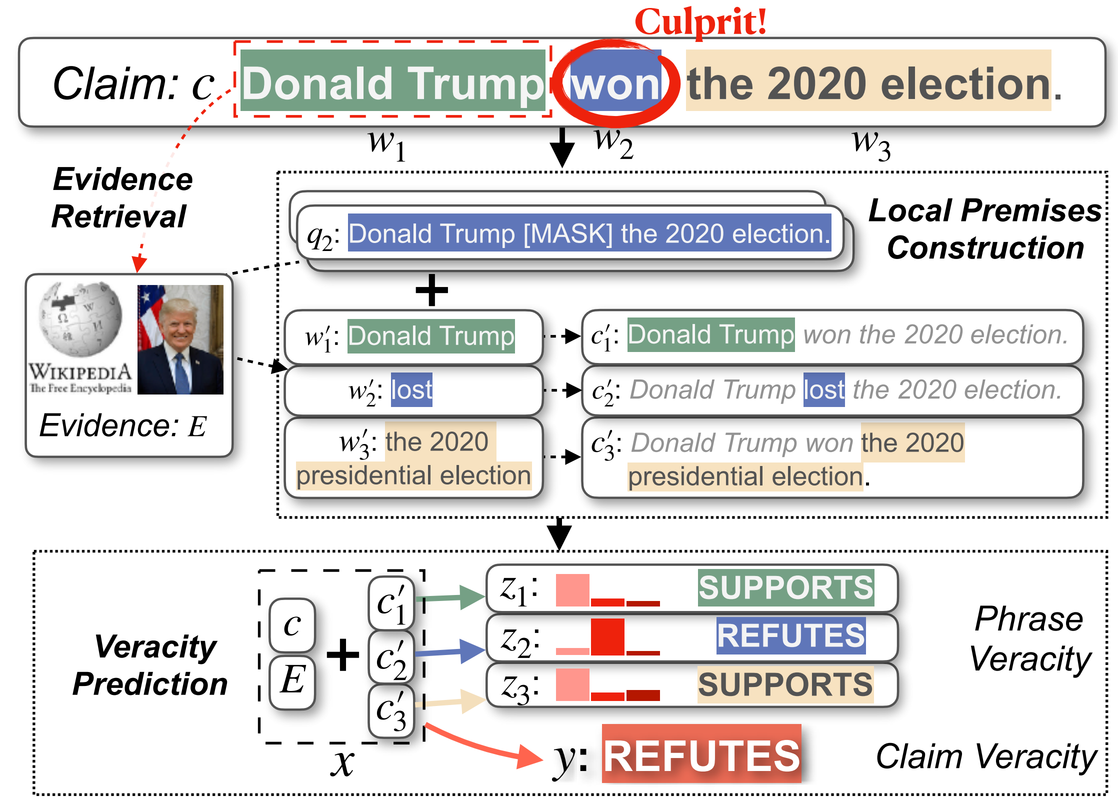
Contribution

❖We propose LOREN for interpretable fact verification.

❖To solve **The Research Problem**, we decompose claim verification at phrase-level, and **build local premises** from evidence to support phrase veracity prediction, **regularized by logical rules**.

The LOREN Framework

Symbolic AI plans, connectionist AI executes.



❖**How to build local premises for verifying claim phrases? — The MRC Solution**

1. Claim Phrase Extraction
2. Probing Question Generation
3. Machine Reading Comprehension

❖**How to train phrase verification without supervision? — The Latent Model**

1. Decompose claim verification $p_\theta(\mathbf{y}\,|\,x)$ into phrase verification $p_\theta(\mathbf{y}\,|\,\mathbf{z},x)$

$$p_\theta(\mathbf{y}\,|\,x) = \sum_{\mathbf{z}} p_\theta(\mathbf{y}\,|\,\mathbf{z},x)p(\mathbf{z}\,|\,x)$$

2. Variational Inference *to solve the latent model*

$$\text{ELBO} = \mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{y},x)}[\log \boxed{p_\theta(y^*\,|\,\mathbf{z},x)}] + D_{\text{KL}}(\boxed{q_\phi(\mathbf{z}\,|\,\mathbf{y},x)}\,\|\,p(\mathbf{z}\,|\,x))$$

❖**How to give latent variables the meaning of phrase veracity? — Regularize Latent Variables with Logic**

$$\mathscr{L}_{\text{logic}}(\theta,\phi) = D_{\text{KL}}\left(p_\theta(\mathbf{y}\,|\,\mathbf{z},x)\;\|\;\boxed{q_\phi^{\text{T}}(\mathbf{y}_z\,|\,\mathbf{y},x)}\right)$$

Soft logic

$$q_\phi^{\text{T}}(\boldsymbol{y}_z = \texttt{SUP}) = \prod_{i=1}^{|\mathbf{z}|} q_\phi(\boldsymbol{z}_i = \texttt{SUP})$$

$$q_\phi^{\text{T}}(\boldsymbol{y}_z = \texttt{REF}) = 1 - \prod_{i=1}^{|\mathbf{z}|}(1 - q_\phi(\boldsymbol{z}_i = \texttt{REF}))$$

$$q_\phi^{\text{T}}(\boldsymbol{y}_z = \texttt{NEI}) = 1 - q_\phi^{\text{T}}(\boldsymbol{y}_z = \texttt{SUP}) - q_\phi^{\text{T}}(\boldsymbol{y}_z = \texttt{REF})$$

❖**Total Loss: -ELBO + Regularization**

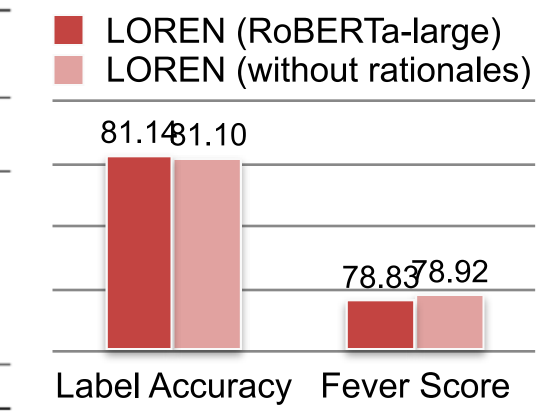
$$\mathscr{L}_{\text{final}}(\theta,\phi) = (1-\lambda)\mathscr{L}_{\text{var}}(\theta,\phi) + \lambda\mathscr{L}_{\text{logic}}(\theta,\phi)$$

Experiments

❖**RQ1: Can we find rationales without hurting verification performance?**

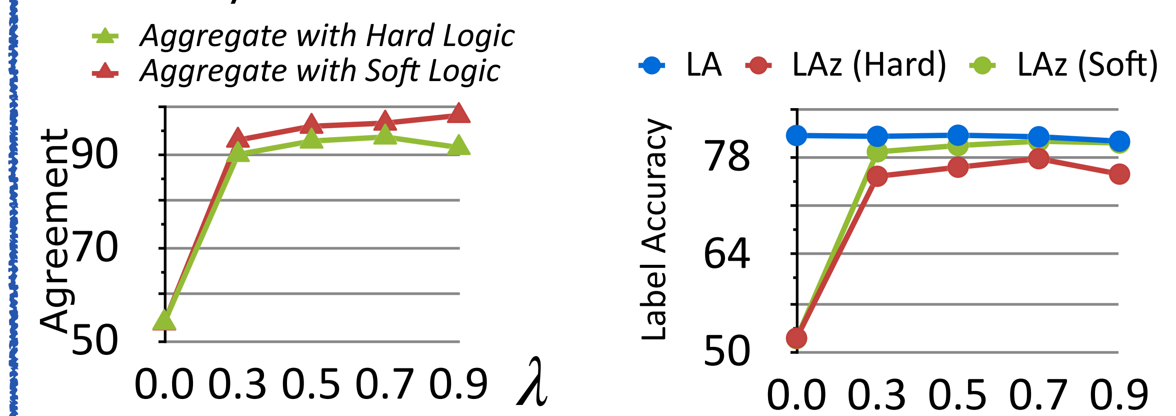
A1: Yes.

Model	Dev LA	Dev FEV	Test LA	Test FEV
UNC NLP	69.72	66.49	68.21	64.21
GEAR ($\text{BERT}_{\text{base}}$)	74.84	70.69	71.60	67.10
DREAM ($\text{XLNet}_{\text{large}}$)	79.16	-	76.85	70.60
KGAT ($\text{BERT}_{\text{large}}$)	77.91	75.86	73.61	70.24
⌞ ($\text{RoBERTa}_{\text{large}}$)	78.29	76.11	74.07	70.38
⌞ ($\text{CorefRoBERTa}_{\text{l}}$)	-	-	75.96	72.30
LOREN ($\text{BERT}_{\text{large}}$)	78.44	76.21	74.43	70.71
⌞ ($\text{RoBERTa}_{\text{large}}$)	81.14	**78.83**	76.42	72.93
LisT5 (T5_{3B})	**81.26**	77.75	**79.35**	**75.87**



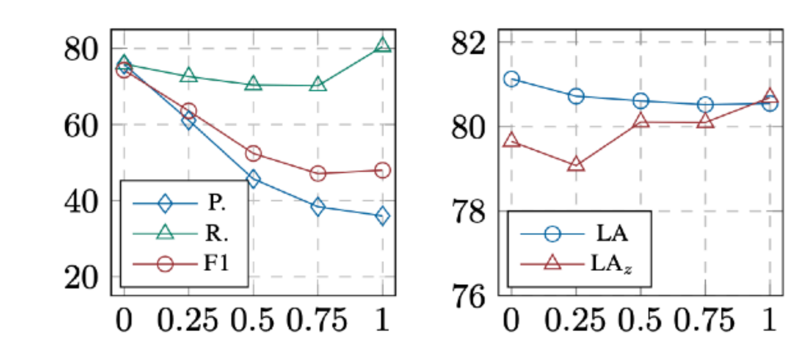
❖**RQ2: How faithful and accurate are these unsupervised rationales?**

A2: Very faithful and accurate.



❖**RQ3: How do local premises contribute to LOREN and its rationales?**

A3: Critical to the quality of phrase veracity prediction.



(a) Results on CULPA. (b) Results on LA & LA_z.

More details in the paper!

Debugging LOREN

Claim2: Ashley Cole is Iranian.
Evidence: Ashley Cole (born 20 December 1980) is an English professional footballer who … in Major League Soccer. Born in Stepney , London…

Premise1: Ashley Cole is Iranian.
Veracity: ***SUPPORTS*** z_1 = [0.981, 0.004, 0.015]
Premise2: Ashley Cole is European.
Veracity: ***REFUTES*** z_2 = [0.014, 0.520, 0.466]

Prediction y**:** REFUTES ✗ y_z = [0.014, 0.522, 0.464] ✗
Ground Truth: NOT ENOUGH INFO

If you have any questions, please email: jjchen19@fudan.edu.cn