The 39th Annual AAAI Conference on Artificial Intelligence

BiDeV: Bilateral Defusing Verification for Complex Claim Fact-Checking

Gaoling School of Artificial Intelligence
Renmin University of China

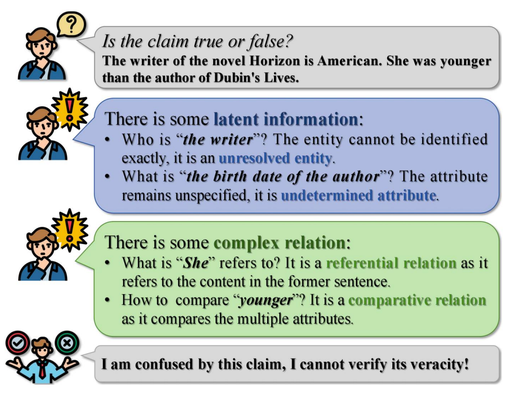
Yuxuan Liu[1], Hongda Sun[1], Wenya Guo[2], Xinyan Xiao[3], Cunli Mao[4], Zhengtao Yu[4], Rui Yan[1*]
[1] Gaoling School of Artificial Intelligence, Renmin University of China
[2] Nankai University [3] Baidu Inc. [4] Kunming University of Science and Technology
yuxuanliu@ruc.edu.cn

X WeChat Code Paper

Introduction

Motivation

- **Claim vagueness makes the claim difficult to understand intuitively.**
- **Evidence redundancy leads to distraction from irrelevant information.**



Charles Russell directed a 1994 superhero fantasy film. The host of the show VJ Logan won had a part in it. Supports

FOLK

Predicates:
Directed(Charles Russell, a 1994 superhero fantasy film) :::
Charles Russell directed a 1994 superhero fantasy film
Host(VJ Logan, a show) :::
VJ Logan host a show
Part(VJ Logan, a 1994 superhero fantasy film) :::
VJ Logan had a part in the 1994 superhero fantasy film.
Followup Question:
What film did Charles Russell direct in 1994? → *The Mask*
What show did VJ Logan host? → *Unknown!*
Did VJ Logan have a part in the 1994 superhero fantasy film? → *No!*
Predicted Label: Refutes ×

ProgramFC

```
def program():
  answer_1 = Question("Who directed the 1994 superhero fantasy film?")
                Charles Russell
  fact_1 = Verify(f"Charles Russell directed the film {answer_1}.")
                False
  answer_2 = Question("Who won the show VJ Logan?")
                VJ Logan
  fact_2 = Verify(f"{answer_2} had a part in the film.")
                False
  label = Predict(fact_1 and fact_2)
                False
```

Predicted Label: Refutes ×

BiDeV

Perceived question in iter-1:
What is the 1994 superhero fantasy film directed by Charles Russell? → *The Mask*
Rewrited claim in iter-1:
Charles Russell directed *The Mask*, a 1994 superhero fantasy film. The host of the show VJ Logan won had a part in *The Mask*.

Perceived question in iter-2:
What is the show that VJ Logan won? → *America's Most Smartest Model*
Rewrited claim in iter-2:
Charles Russell directed The Mask, a 1994 superhero fantasy film. The host of *America's Most Smartest Model* which VJ Logan won had a part in The Mask.

Perceived question in iter-3:
Who is the host of the show VJ Logan won? → *Ben Stein*
Rewrited claim in iter-3:
Charles Russell directed The Mask, a 1994 superhero fantasy film, *Ben Stein*, the host of America's Most Smartest Model which VJ Logan won, had a part in The Mask.

Decomposed Sub-claims:
Charles Russell directed The Mask, a 1994 superhero fantasy film. → *True*
Ben Stein had a part in The Mask. → *True*
Ben Stein was the host of America's Most Smartest Model. → *True*
VJ Logan won America's Most Smartest Model. → *True*

Predicted Label: Supports √

Contribution

- We propose BiDeV, **a novel fact-checking framework** integrating LLMs to eliminate vague information in the claim and noisy redundancy in the evidence.
- We introduce the **Vagueness Defusing** for claim simplification, which focuses on ascertaining latent information and resolving complex relations.
- We present the **Redundancy Defusing** for evidence selection, which filteres out irrelevant information leading to more effective and pertinent evidence.

Framework Overview

BiDeV: Bilateral Defusing Verification

- **Vagueness Defusing focuses on claim simplification.**
- **Redundancy Defusing focuses on evidence selection.**

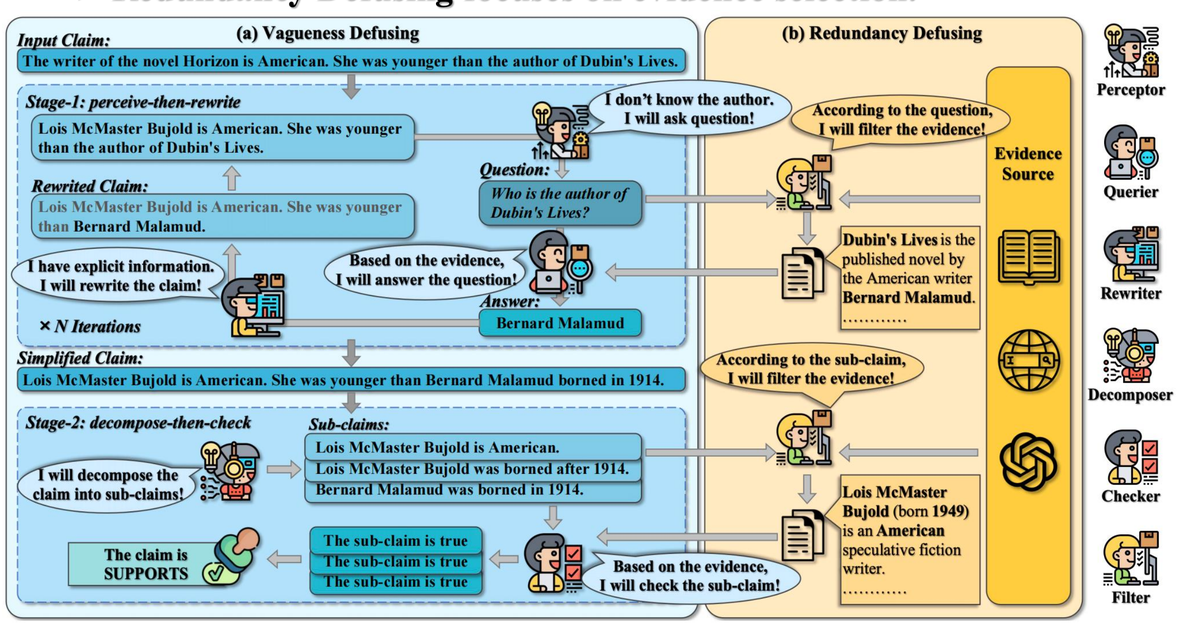


Figure 2: The overview of our BiDeV. Two main modules for Bilateral Defusing Verification: (a) **Vagueness Defusing** for input claim. *Perceive-then-rewrite* stage simplifies the claim iteratively: the perceptor perceives questions about latent information, the querier provides explicit knowledge to the question and the rewriter rewrites the latent information in the claim with the explicit knowledge. *Decompose-then-check* stage verifies the claim: the decomposer splits several sub-claims and the checker verifies the sub-claims. (b) **Redundancy Defusing** for evidence. The evidence extracted from the source is refined by the filter.

Bilateral Defusing Verification

Vagueness Defusing

Stage 1: Perceive-then-Rewrite for Latent Information

Two types of Latent Information:

- ***Unresolved Entities**: entity referred to cannot be found in the claim.*
- ***Undetermined Attributes**: attribute required is not mentioned in the claim*

Perceptor identifies latent information and generates targeted question for obtaining explicit background knowledge.

$$q_i = M_p(c_{i-1})$$

Querier answers the question generated by the perceptor with the precise and explicit content of the latent information.

$$a_i = M_q(q_i, e_i^*)$$

Rewriter rewrites the direct counterparts and the indirect relevance of the generated questions with the obtained answer.

$$c_i = M_r(c_{i-1}, q_i, a_i)$$

Stage 2: Decompose-then-Check for Complex Relations

Two types of Complex Relations:

- ***Referential Relation**: refers to some content in the former sentence.*
- ***Comparative Relation**: compares multiple attributes in the sentence.*

Decomposer replaces referential relation with explicit entities in the claim and splits comparative relation using determined attributes.

$$sc = M_d(C^*)$$

Checker conducts the final step of fact-checking to verify each sub-claims and conclude the veracity result of the entire claim.

$$Y = \bigcap_{j}^{|sc|} y_j, \quad y_j = M_c(sc_j, e_j^*)$$

Redundancy Defusing

Filter segments the extracted evidence into multiple paragraphs and evaluates whether each paragraph is relevant to the question or the sub-claim, then filters out irrelevant paragraphes.

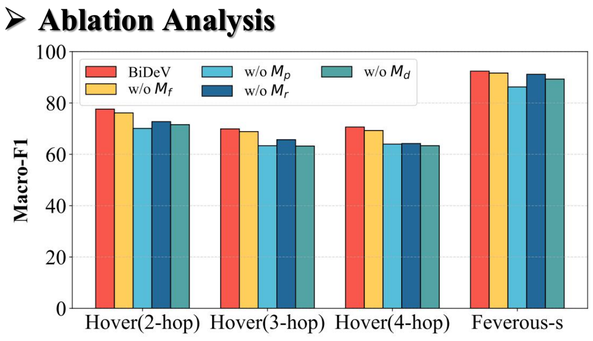
$$e_i^* = M_f(e_i, q_i) \quad e_j^* = M_f(e_j, sc_j)$$

Experimental Results

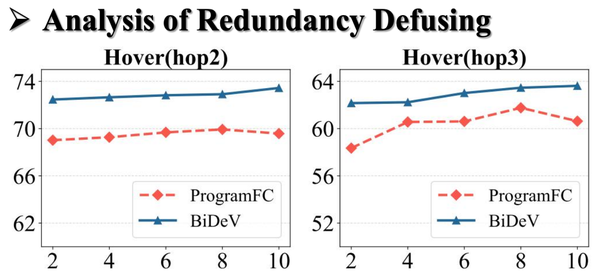
Overall Performance

Methods	Hover(hop2)		Hover(hop3)		Hover(hop4)		Feverous-s	
	Gold	Open	Gold	Open	Gold	Open	Gold	Open
Bert-FC* (Soleimani, Monz, and Worring 2020)	53.41	50.68	50.91	49.86	50.86	48.57	74.71	51.67
LisT5* (Jiang, Pradeep, and Lin 2021)	56.15	52.56	53.76	51.89	51.67	50.46	77.88	54.15
RoBERTa-NLI* (Nie et al. 2020)	74.62	63.62	62.23	53.99	57.98	52.41	88.28	57.81
DeBERTaV3-NLI* (He, Gao, and Chen 2022)	77.22	68.72	65.98	60.76	60.49	56.01	91.98	58.81
MULTIVERS* (Wadden et al. 2022)	68.86	60.17	59.87	52.55	55.67	51.86	86.03	56.61
Codex* (Chen et al. 2021)	70.63	65.07	66.46	56.63	63.49	57.27	89.77	62.58
FLAN-T5* (Chung et al. 2024)	73.69	69.02	65.66	60.23	58.08	55.42	90.81	63.73
HiSS† (Zhang and Gao 2023)	73.06	66.25	65.14	58.56	64.67	57.64	89.26	65.99
FOLK† (Wang and Shu 2023)	73.24	67.29	65.84	58.61	64.73	58.79	89.52	66.89
ProgramFC† (Pan et al. 2023)	74.59	69.89	66.75	61.21	65.00	58.21	91.23	67.22
Factcheck-GPT† (Wang et al. 2024)	74.88	70.25	66.32	60.11	66.62	59.25	91.39	67.24
BiDeV	**77.59**	**73.44**	**69.91**	**63.62**	**70.63**	**60.41**	**92.39**	**69.01**

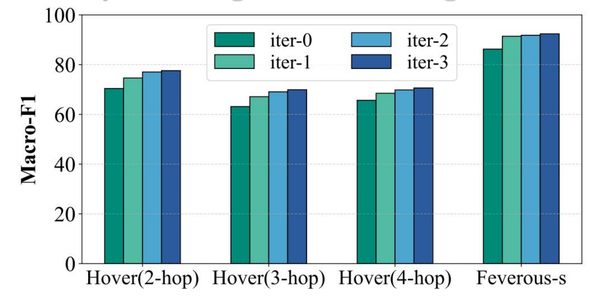
Ablation Analysis



Analysis of Redundancy Defusing



Analysis of Vagueness Defusing



Analysis of Model Scale in Querier and Checker