MultiVerS

Improving scientific claim verification with weak supervision and full-document context

CS221.O12.KHCL

Instructor: PhD. Nguyen Thi Quy

Group 15

Le Gia Khang Nguyen Hoang Tan Le Duy Khang

December 17, 2023

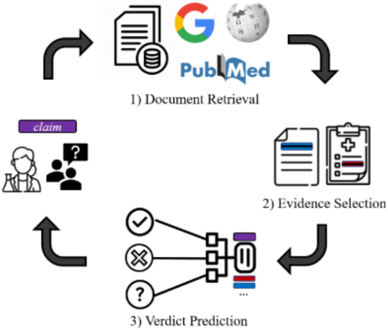
Task

Definition of scientific claim verification from the SCIFACT task:

Given a claim c and a *candidate abstract* a.

Label $y(c, a) \in \{SUPPORTS, REFUTES, NEI\}$.

Identify rationales $R(c, a) = \{r_1(c, a), ..., r_n(c, a)\}$.




```
{
  "id": 78526,
  "predicted label": "REFUTES",
  "predicted evidence": [
    [
      "Lorelai Gilmore",
      3
    ]
  ]
}
```

Task

Claim:

Advil (ibuprofen) worsens COVID-19 symptoms

Evidence abstract:

Covid-19 and avoiding Ibuprofen.

...

Increased risk of COVID-19 infection was feared with ibuprofen use

...

At this time, there are no findings discouraging the use of ibuprofen

Task

Claim:

Advil (ibuprofen) worsens COVID-19 symptoms

Label: Refuted

Evidence abstract:

Covid-19 and avoiding Ibuprofen.

...

Increased risk of COVID-19 infection was feared with ibuprofen use

...

At this time, there are no findings discouraging the use of ibuprofen

Task

Claim:

Advil (ibuprofen) worsens COVID-19 symptoms

Label: Refuted

Task Outputs

1. Fact-checking label

Evidence abstract:

Covid-19 and avoiding Ibuprofen.

...

Increased risk of COVID-19 infection was feared with ibuprofen use

...

At this time, there are no findings discouraging the use of ibuprofen

Task

Claim:

Advil (ibuprofen) worsens COVID-19 symptoms

Label: Refuted

Task Outputs

1. Fact-checking label
2. Rationales justifying the label

Evidence abstract:

Covid-19 and avoiding Ibuprofen.

...

Increased risk of COVID-19 infection was feared with ibuprofen use

...

At this time, there are no findings discouraging the use of ibuprofen

Rationale

Task

Claim:

Advil (ibuprofen) worsens COVID-19 symptoms

Label: Refuted

Task Outputs

1. Fact-checking label
2. Rationales justifying the label

Evidence abstract:

Covid-19 and avoiding Ibuprofen.

...

Increased risk of COVID-19 infection was feared with ibuprofen use

Context required

...

At this time, there are no findings discouraging the use of ibuprofen

Rationale

Prior work: Extract-then-label

Claim:

Advil (ibuprofen) worsens COVID-19 symptoms

Evidence abstract:

Covid-19 and avoiding Ibuprofen.

...

Increased risk of COVID-19 infection was feared with ibuprofen use

...

At this time, there are no findings discouraging the use of ibuprofen

Prior work: Extract-then-label

Claim:

Advil (ibuprofen) worsens COVID-19 symptoms

At this time, there are no findings discouraging the use of ibuprofen

Evidence abstract:

Covid-19 and avoiding Ibuprofen.

...

Increased risk of COVID-19 infection was feared with ibuprofen use

...

At this time, there are no findings discouraging the use of ibuprofen

Prior work: Extract-then-label

Claim:

Advil (ibuprofen) worsens COVID-19 symptoms

At this time, there are no findings discouraging the use of ibuprofen

Label: Refuted

Evidence abstract:

Covid-19 and avoiding Ibuprofen.

...

Increased risk of COVID-19 infection was feared with ibuprofen use

...

At this time, there are no findings discouraging the use of ibuprofen

Prior work: Extract-then-label

Claim:

Advil (ibuprofen) worsens COVID-19 symptoms

At this time, there are no findings discouraging the use of ibuprofen

Label: Refuted

Evidence abstract:

Covid-19 and avoiding Ibuprofen.

...

Increased risk of COVID-19 infection was feared with ibuprofen use

...

At this time, there are no findings discouraging the use of ibuprofen

Drawbacks of extract-then-label:

1. Rationales may lack context
2. Requires rationale supervision during training

Prior models

These models make predictions in 2 steps:

Predict rationales $\hat{R}(c,a) = \{\hat{r}_1(c,a), ...\hat{r}_n(c,a)\}$

Then, make a label prediction $\hat{y}(c, f_R(\hat{R}(c,a)))$

Given a claim c and candidate abstract a

MultiVerS

A multitask system for full-context scientific claim verification

- Predict $\hat{y}(c, a)$ directly based on an encoding of the entire claim and abstract.
- Enforce consistency of $\hat{R}(c, a)$ with $\hat{y}(c; a)$ during decoding.

MultiVerS

Long document encoding:

A claim c and candidate abstract a consisting of title t and sentences $s_1, \ldots, s_n$.

The $< /s >$ token following each sentence s_i is notated as $< /s >_i$.

$$\text{<s>}\ c\ \text{</s>}\ t\ \text{</s>}\ s_1\ \text{</s>}_1\ \ldots s_n\ \text{</s>}_n$$

Global attention is assigned to $< s >$ token, all tokens in c and all $< /s >$ tokens.

MultiVerS

Claim

Advil (ibuprofen) worsens COVID-19 symptoms

Abstract sentence 1

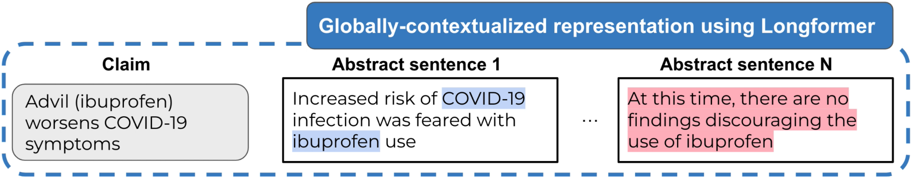
Increased risk of COVID-19 infection was feared with ibuprofen use

...

Abstract sentence N

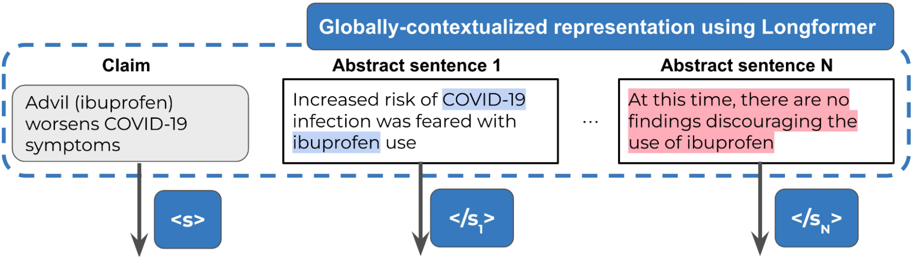
At this time, there are no findings discouraging the use of ibuprofen

MultiVerS



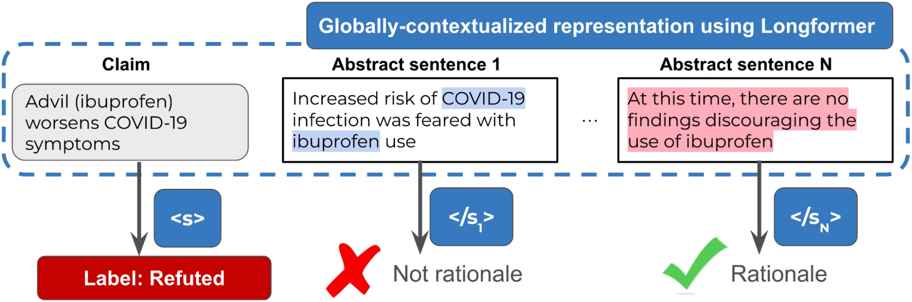
Globally-contextualized representation using Longformer
Claim
Advil (ibuprofen) worsens COVID-19 symptoms
Abstract sentence 1
Increased risk of COVID-19 infection was feared with ibuprofen use
...
Abstract sentence N
At this time, there are no findings discouraging the use of ibuprofen

MultiVerS



Rationale Prediction: The model analyzes each sentence's ending $< /s >$ to decide if it justifies the claim, using a binary classification.

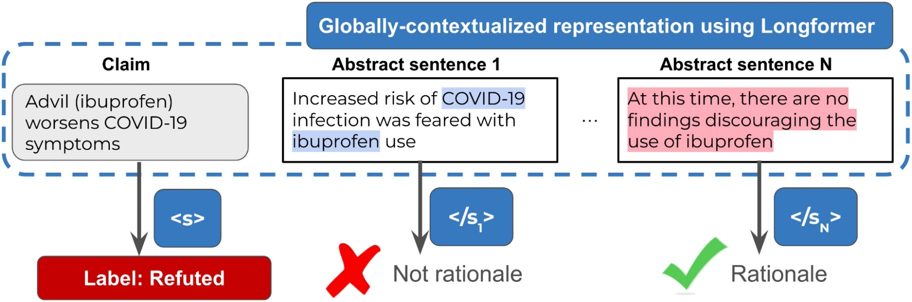
MultiVerS



Label Prediction: A special $< s >$ token captures the essence of both the claim and the abstract through "global attention".

$\longrightarrow$ Enabling accurate fact-checking label prediction for the abstract.

MultiVerS



$$\mathcal{L} = \mathcal{L}_{label} + \lambda_{rationale}\mathcal{L}_{rationale}$$

Benefits of multitask approach:

1. Incorporates all relevant context
2. Can train on instances with no rationale annotations

Experiments

Dataset	**Domain**	**Claim source**	**Open**	**Has NEI**	**Claim complexity**	**Negation method**	**Train claims**	**Eval claims**	>512 **tokens**
HealthVer	COVID	TREC-COVID	✗	✓	Complex	Natural	1,622	230	14.9%
COVIDFact	COVID	Reddit	✗	✗	Complex	Automatic	903	313	12.4%
SCIFACT	Biomed	Citations	✓	✓	Atomic	Human	1,109	300	27.4%
FEVER	Wiki	Wikipedia	-	✓	Atomic	Human	130,644	-	33.2%
PUBMEDQA	Biomed	Paper titles	-	✓	Complex	Automatic	58,370	-	12.1%
EVIDENCEINFERENCE	Biomed	ICO prompts	-	✓	Atomic	Automatic	7,395	-	42.7%

Table: Summary of datasets used in experiments

Experiments

Target datasets:

- HealthVer
- COVID-Fact
- SciFact

Experiments

Target datasets:

- HealthVer
- COVID-Fact
- SciFact

Roughly 1000 claims / dataset.

Expert annotations are expensive

Experiments

Target datasets:

- HealthVer
- COVID-Fact
- SciFact

Roughly 1000 claims / dataset.

Expert annotations are expensive

Traning procedure:

- **Stage 1:** Train on a combination of *labeled out of domain* data *weakly-labeled in-domain data.*
- **Stage 2:** Continue training on data from each target dataset.

Experiments

Target datasets:

- HealthVer
- COVID-Fact
- SciFact

Roughly 1000 claims / dataset.

Expert annotations are expensive

Traning procedure:

- **Stage 1:** Train on a combination of *labeled out of domain* data *weakly-labeled in-domain data.*
- **Stage 2:** Continue training on data from each target dataset.

Domain adaptation settings:

- **Zero-shot:** Stage 1 training only.
- **Few-shot:** 45 instances from target datasets.
- **Full-supervised:** All target data.

Data: Stage 1

Supervised out-of-domain data (FEVER)

LeBron James was born in Ohio

Label: Supported

LeBron James is an American basketball player. He was born in Akron, Ohio.

Data: Stage 1

Supervised out-of-domain data (FEVER)

LeBron James was born in Ohio

Label: Supported

LeBron James is an American basketball player. He was born in Akron, Ohio.

Weakly-supervised in-domain data

Diabetes increases risk of depression

Abstract

...

...

...

Label: Supported

Data: Stage 1

Supervised out-of-domain data (FEVER)

LeBron James was born in Ohio

Label: Supported

LeBron James is an American basketball player. He was born in Akron, Ohio.

Weakly-supervised in-domain data

Diabetes increases risk of depression

Abstract

...

...

...

Label: Supported

Claim: Paper title

Data: Stage 1

Supervised out-of-domain data (FEVER)

LeBron James was born in Ohio

Label: Supported

LeBron James is an American basketball player. He was born in Akron, Ohio.

Weakly-supervised in-domain data

Data: Stage 1

Supervised out-of-domain data (FEVER)

LeBron James was born in Ohio

Label: Supported

LeBron James is an American basketball player. He was born in Akron, Ohio.

Weakly-supervised in-domain data

MultiVerS can train on these examples, even though no rationale annotations are provided

Evaluation

Abstract-level evaluation:

- Identifying abstracts that SUPPORT or REFUTE each claim.
- Predicting the correct label $y(c, a)$ is sufficient

Sentence-level evaluation:

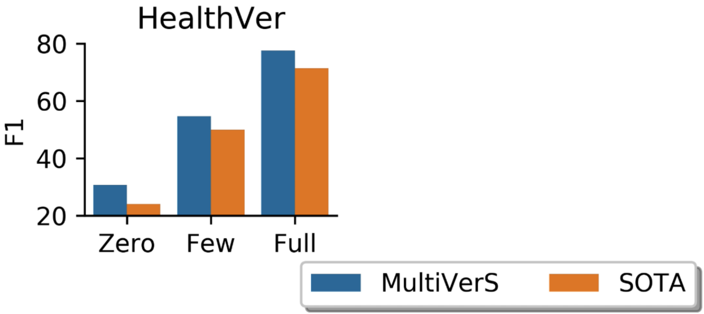
- It combines the accuracy of abstract-level label prediction with the precision of rationale identification.

Evaluation

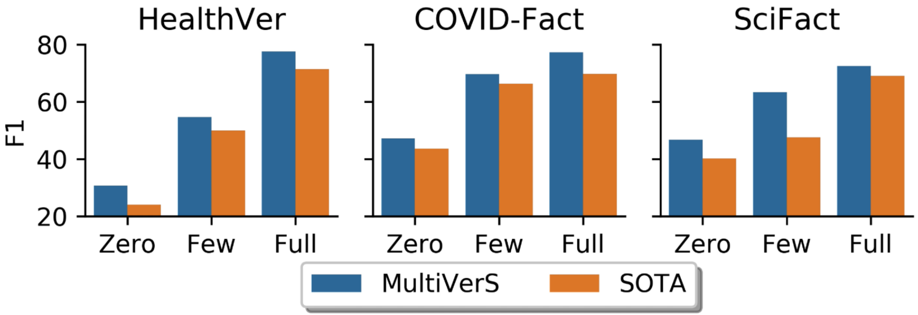
Setting	Model	HealthVer Abstract P	R	F1	HealthVer Sentence P	R	F1	COVIDFact Abstract P	R	F1	COVIDFact Sentence P	R	F1	SciFact Abstract P	R	F1	SciFact Sentence P	R	F1
Zero	ParagraphJoint	72.3	14.4	24.0	22.9	2.7	4.9	51.3	37.9	43.6	31.5	16.0	21.3	52.9	32.4	40.2	36.4	14.9	21.1
	MultiVerS	60.6	20.5	**30.7**	25.0	4.6	**7.8**	48.8	45.7	**47.2**	32.7	18.5	**23.6**	49.0	44.6	**46.7**	39.0	21.6	**27.8**
Few	ParagraphJoint	62.7	41.6	50.0	46.0	29.3	**35.8**	73.3	60.6	66.3	44.3	30.6	36.2	44.4	51.4	47.6	33.0	35.1	34.0
	MultiVerS	63.6	47.9	**54.7**	41.9	31.0	35.7	71.3	68.1	**69.7**	39.5	35.4	**37.4**	76.4	54.1	**63.3**	51.7	40.3	**45.3**
Full	Vert5Erini	71.3	74.0	72.6	65.6	61.2	63.3	76.6	52.7	62.4	44.8	27.2	33.9	64.0	73.0	68.2	60.6	66.5	63.4
	ParagraphJoint	75.0	68.3	71.5	69.9	60.6	64.9	71.5	68.1	69.8	41.4	40.3	40.8	75.8	63.5	69.1	68.9	54.6	60.9
	MultiVerS	78.9	76.3	**77.6**	71.4	67.0	**69.1**	77.3	77.3	**77.3**	41.5	46.1	**43.7**	73.8	71.2	**72.5**	67.4	67.0	**67.2**

Table 2: Performance of MultiVerS and baselines.

Results

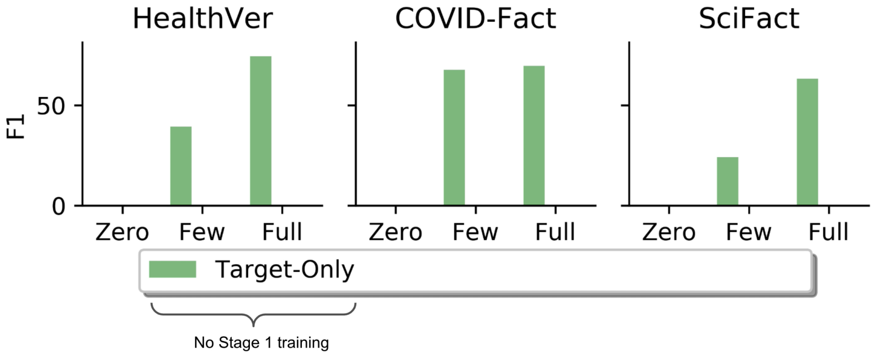


Results

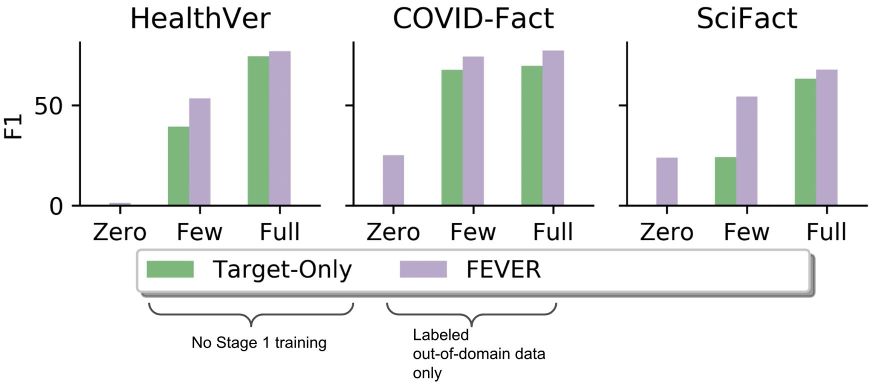


MultiVerS outperforms SOTA on all datasets

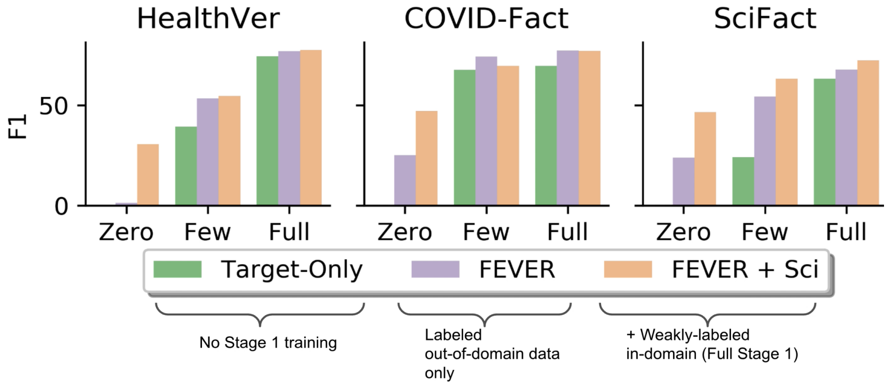
Ablations: Training strategy



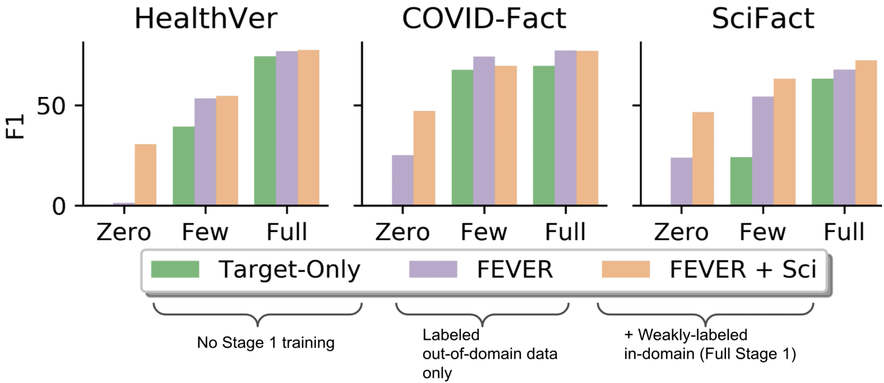
Ablations: Training strategy



Ablations: Training strategy



Ablations: Training strategy



Pretraining with weakly-supervised in-domain data improves few / zero shot performance.

Reference

MULTIVERS: Improving scientific claim verification with weak supervision and full-document context

Scientific Fact-Checking: A Survey of Resources and Approaches
Juraj Vladika and Florian Matthes

Longformer: The Long-Document Transformer
Iz Beltagy, Matthew E. Peters and Arman Cohan

Code and model checkpoints for the MultiVerS model
dwadden/multivers

Thanks for listening!

Q&A section