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
- Factualy accurate language generation

Interpretable Fact Verification
- Right answer for 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?

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 into phrase verification

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

Total Loss: -ELBO + Regularization

LOREM and its rationales?

RQ1: Can we find rationales without hurting verification performance? A1: Yes.

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.

Debugging LOREN

LOREN (without rationales)

LOREN (RoBERTa-large)

More details in the paper!