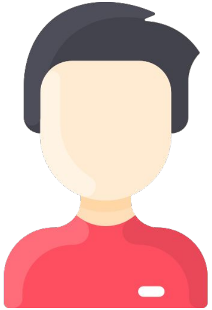
Interpreting Indirect Answers Using Self-Rationalizing Models

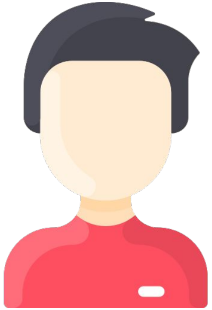
Yun Li, Michael Neely, Frederik Nolte

When asked a polar question...



Question: Do you like red meat?

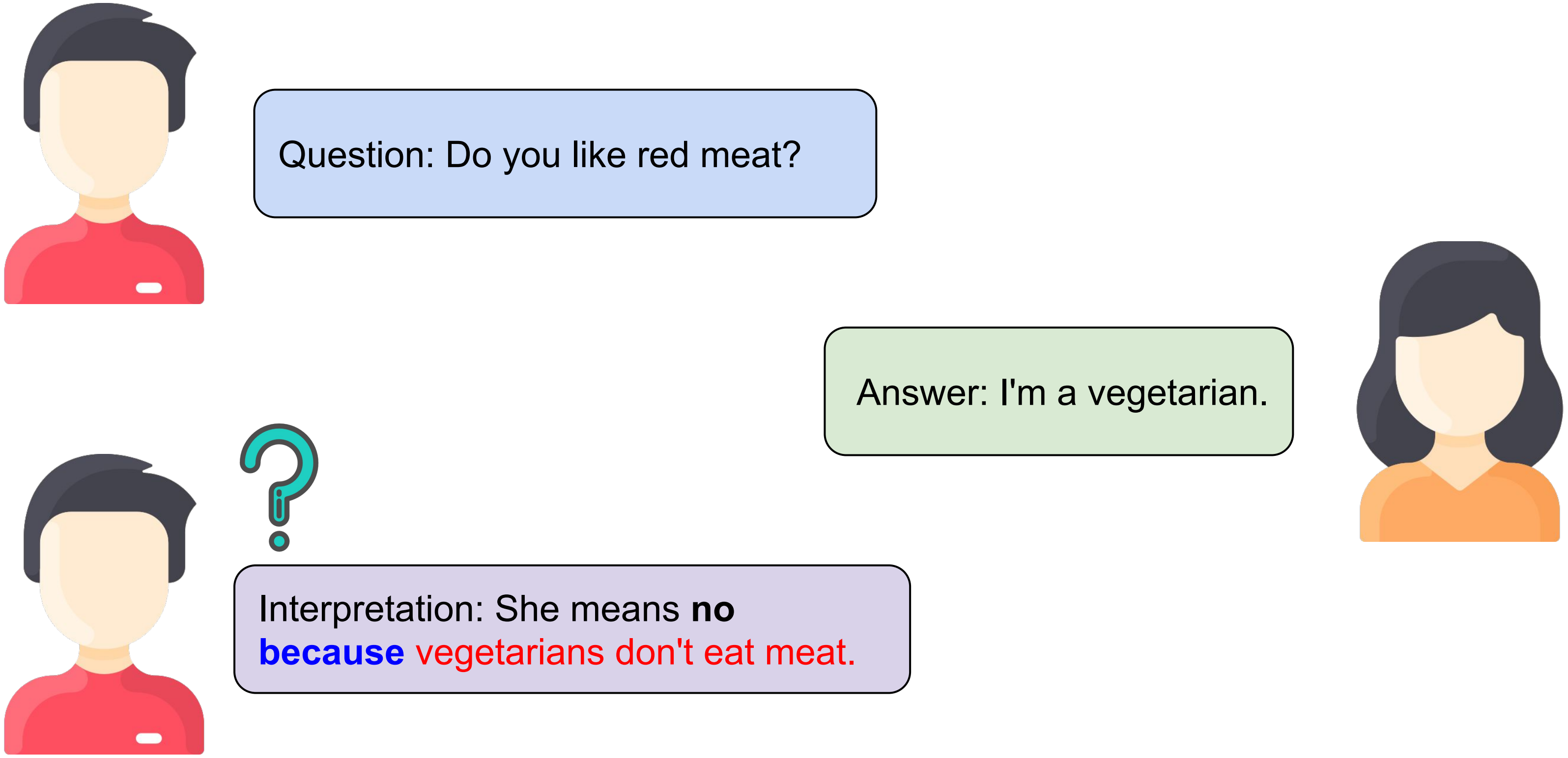
...People can respond indirectly



Question: Do you like red meat?

Answer: I'm a vegetarian.

Humans intuitively interpret such responses



But what about modern neural language models?

- Circa Dataset (Louis et al., 2020) to test interpretation capacity.
  - 34,268 crowdsourced (context, polar question, indirect answer) pairs
  - Two label settings:

Label	STRICT	
Yes	14,504	(42.3%)
No	10,829	(31.6%)
Probably yes / sometimes yes	1,244	(3.6%)
Yes, subject to some conditions	2,583	(7.5%)
Probably no	1,160	(3.4%)
In the middle, neither yes nor no	638	(1.9%)
I am not sure	63	(0.2%)
Other	504	(1.5%)
N/A	2,743	(8.0%)

Table 7: Distribution of STRICT gold standard labels. 'N/A' indicates lack of majority agreement.

Our focus

Label	RELAXED	
Yes	16,628	(48.5%)
No	12,833	(37.5%)
Yes, subject to some conditions	2,583	(7.5%)
In the middle, neither yes nor no	949	(2.8%)
Other	504	(1.5%)
N/A	771	(2.2%)

Table 8: Distribution of RELAXED gold standard labels. 'N/A' indicates lack of majority agreement.

But what about modern neural language models?

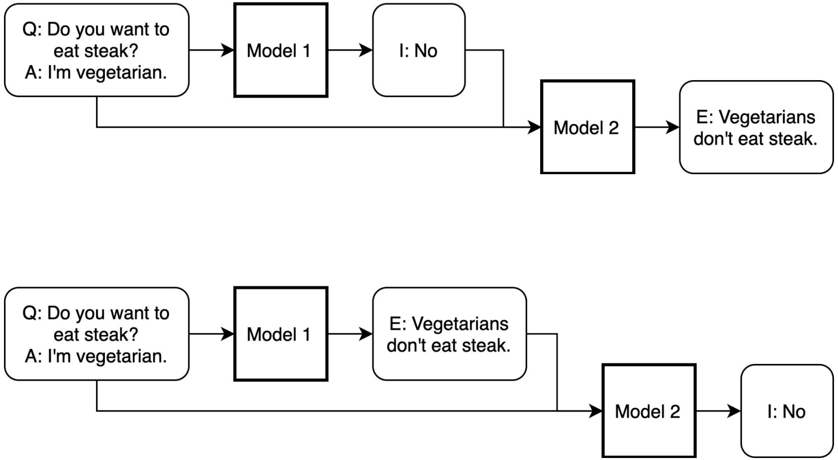
1. Expectation: It will be **difficult**, because interpreting indirect answers requires extensive amounts of *background knowledge* and *common sense*
2. Reality: **easy** to classify
    - Multiple Choice QA models can easily reach 90%+ accuracy in the relaxed setting, and can reach 80+% accuracy with the answer only.
    - Why?
        - Annotation artifacts: simple pattern matching and co-occurrence statistics
        - Poor benchmarking task: does not require reading comprehension skills, common sense reasoning, or world knowledge (see e.g., Sugawara et al., 2019)

What if we raise the bar?

- If we ask a model to rationalize (explain) its decision, we can more accurately gauge its ability to understand natural language.
- Faithfulness
    - Accurately represents the reasoning process behind the model's prediction (Jacovi and Goldberg, 2020)
    - If the model predicts the wrong label, we can see why it made a mistake
    - If the model predicts the right label, we can see if makes the correct logical inference
- **Best type of explanation: free-form natural language text** (unrestricted)

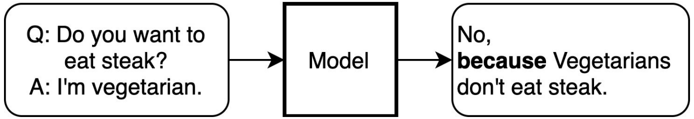
Options

Pipeline: Explain-Then-Predict



e.g., (Latcinnik and Berant, 2020)

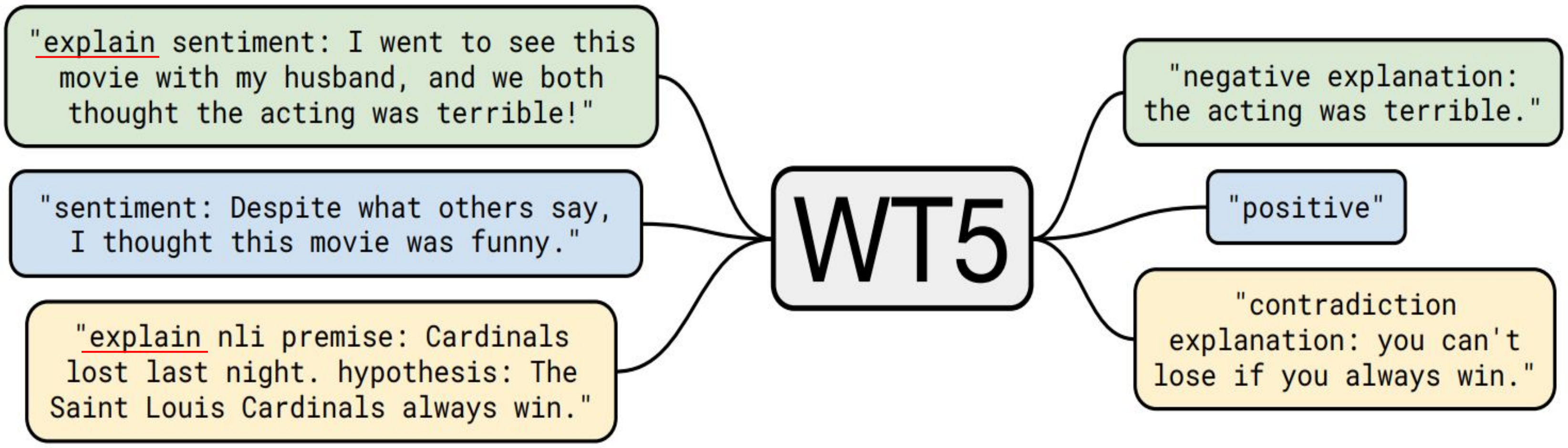
Jointly: Predict and Rationalize



More Faithful (Wiegreffe et al., 2020)

Method

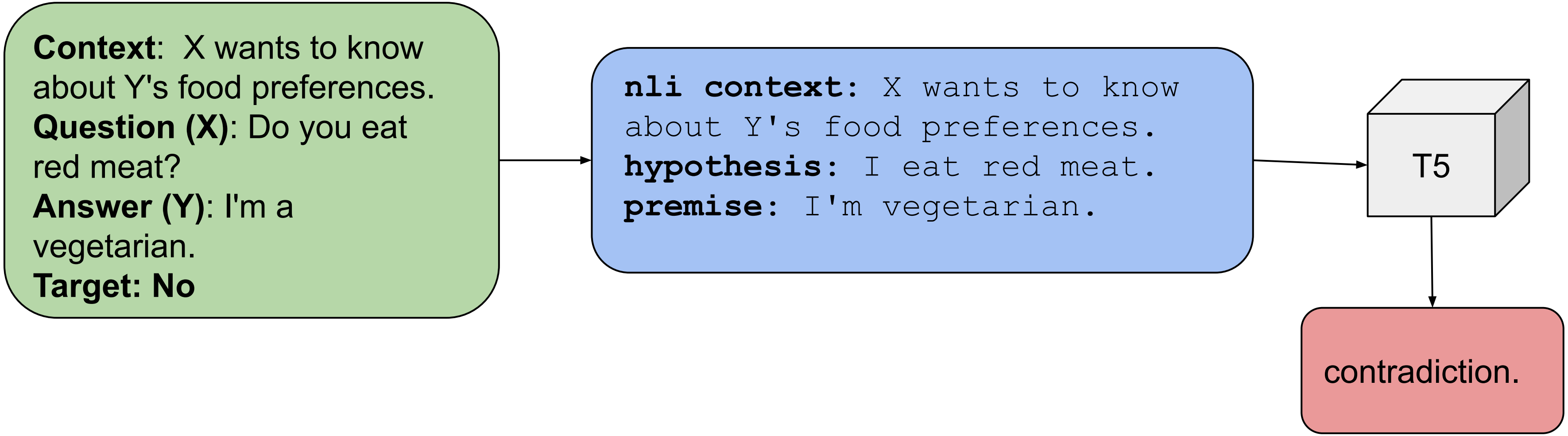
- Text-to-Text with T5 (Raffel et al., 2020)
- Same setup as Narang et al., 2020: "WT5"



Method

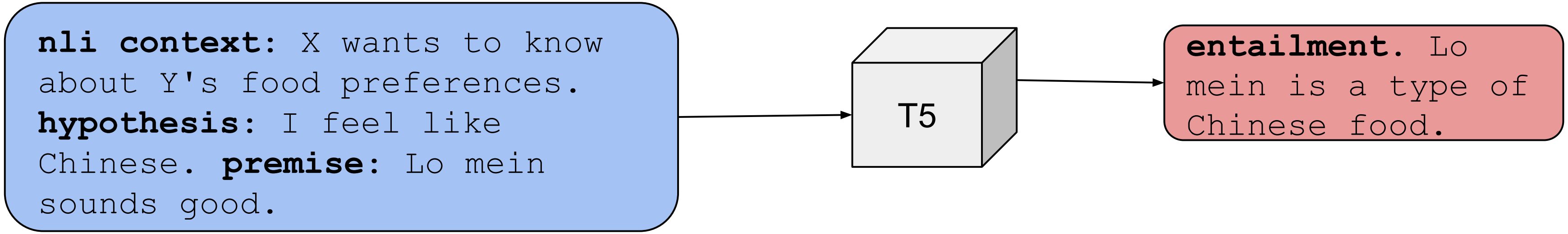
- One problem: no references with which to supervise rationale generation!
- Solution (similar to Narang et al., 2020): transfer learning
    - e-SNLI (Camburu et al., 2018): ~570k Natural language inference (NLI) instances with human-provided explanations
    - CoS-E (Rajani et. al, 2019): ~10k multiple choice questions from CommonSense QA (Talmor et al., 2019) with human-provided explanations
- **Training**: Finetune T5 on 5-task mixture:
    - e-SNLI with/without rationales, CoS-E with/without rationales, Circa without rationales
- **Evaluation**: Predict and rationalize held out Circa test set

Training



Context: X wants to know about Y's food preferences.
Question (X): Do you eat red meat?
Answer (Y): I'm a vegetarian.
Target: No
nli context: X wants to know about Y's food preferences.
hypothesis: I eat red meat.
premise: I'm vegetarian.
T5
contradiction.

Evaluation


nli context: X wants to know about Y's food preferences. **hypothesis:** I feel like Chinese. **premise:** Lo mein sounds good.
T5
entailment. Lo mein is a type of Chinese food.

Evaluation measures

- Predictive power:
  - Accuracy
  - Micro F1
- Rationale quality and faithfulness:
  - Human judgments (Mechanical Turk survey - Likert scale)
  - **Leakage-Adjusted Simulatability** (Hase et al., 2020)
    - Intuition: when do explanations support model behavior?
    - Gauge "*how well an observer can use model explanations to predict the model's output, while controlling for explanation leakage*" (per Hase et al., 2020)
    - Observer in our experiments: DistilBERT (Sanh et al., 2019)

Results

- Highly accurate
  - ~88% accuracy on held-test test set with unseen contexts
- Very faithful
  - LAS score on par with human results reported by Hase et al., 2020
- Not very creative
  - 39% of rationales match 9 verbose logical regexes
  - Another 41% match 3 naive logical regexes
  - The rest are mostly moderate modifications to the some combination of the premise, hypothesis, and/or context
- Displays average capacity for fluency and reasoning
  - Sometimes applies logical templates that do not match the predicted label
  - Most poor rationales would be correct with just one logical flip (e.g., 'would' → 'would not')
  - Average rationale quality rating is 3/5

Discussion

contradiction. Thai food is not the same as Thai food.

entailment. I can't wait to get in my PJs implies I am excited to go home.

- Semi-intelligent pattern matching
  - Applies templates from e-SNLI and CoS-E based on the predicted label of the Circa instance
  - Because the templates are (mostly) logical in nature, annotators tend to rate the rationales very highly or very poorly. E.g., either the rationale is correct or completely wrong
- Rationales tend to leak
  - Classifiers like DistilBERT can reach ~99% accuracy on a large subset of the rationales
- Still not a true test of language understanding, but a step in the right direction
  - Can reasonably transfer rational generating capacity to new datasets, but only when in a similar domain

References I

- Icons: https://www.flaticon.com/
- O.-M. Camburu, T. Rockt¨aschel, T. Lukasiewicz, and P. Blunsom. e-snli: Natural language inference with natural language explanations, 2018.
- A. Jacovi and Y. Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4198–4205, Online, July 2020. Association for Computational Linguistics.
- Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, pages 4351–4367.
- V. Latcinnik and J. Berant. Explaining question answering models through text generation, 2020.
- Annie Louis, Dan Roth, and Filip Radlinski. 2020. "I'd rather just go to bed": Understanding indirect answers.
- Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. WT5?! Training text-to-text models to explain their predictions.

References II

- N. F. Rajani, B. McCann, C. Xiong, and R. Socher. Explain yourself! leveraging language models for commonsense reasoning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4932–4942, Florence, Italy, July 2019. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1–67.
- Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. Assessing the benchmarking capacity of machine reading comprehension datasets. Proceedings of the AAAI Conference on Artificial Intelligence, 34(05):8918–8927, Apr. 2020.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421.
- Sarah Wiegreffe, Ana Marasovic, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales.