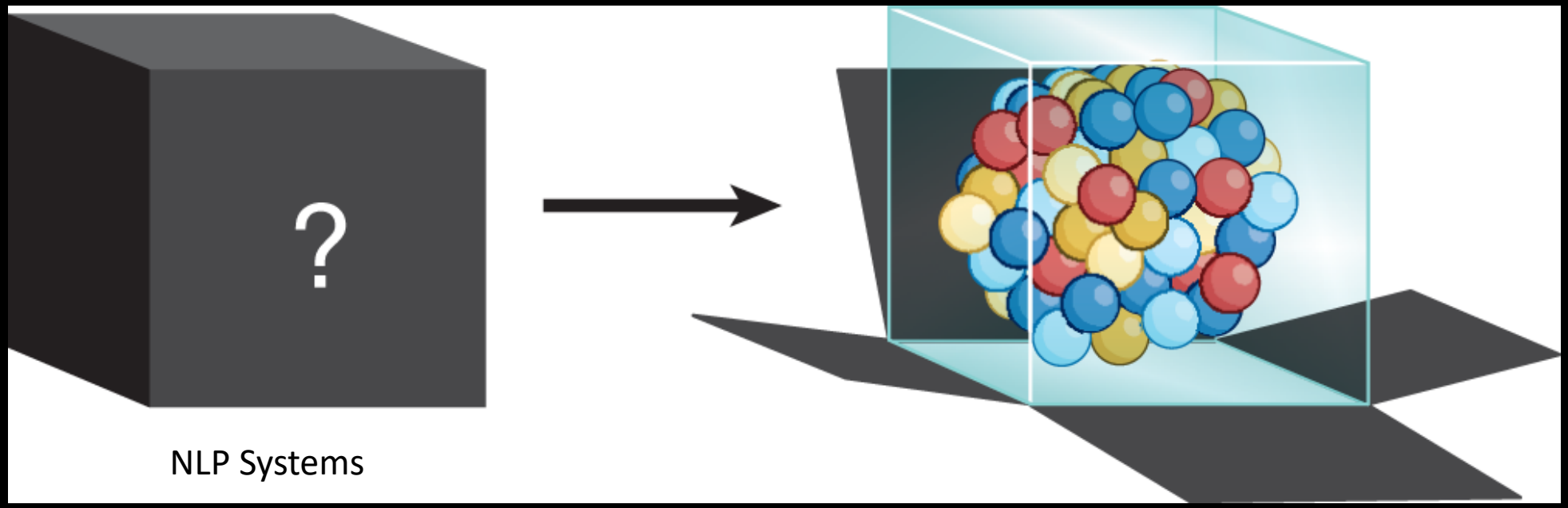


Interpretable NLP

CSC401/2511 – Natural Language Computing – Winter 2023

Lecture 12

University of Toronto

Too complicated to trust

- Transparency
  - How does it work?
  - How can we trust that it works?
- Accountability
  - Why does it *fail*?
  - How can we improve the system?

- People don’t trust a system with poor transparency and accountability.

UNIVERSITY OF TORONTO

Aside: Bill C-27

- In 2022, the Government of Canada tabled Bill C-27.
- This creates new rules for the responsible **development** and **deployment** of AI. Some items include:
  - Increasing **control and transparency** when Canadian's personal information is handled by organizations.
  - Establishing stronger protections for minors, including by limiting organizations' **right to collect or use information** on minors and holding organizations to a higher standard when **handling minors' information**.
  - Many others.

https://ised-isde.canada.ca/site/innovation-better-canada/en/canadas-digital-charter/bill-summary-digital-charter-implementation-act-2020

UNIVERSITY OF TORONTO

Interpretability in NLP

These terminologies are floating everywhere, for example:

- Transparency
- Accountability
- Explainability, explainable AI, XAI
- Interpretability

Interpretability research involves everything that makes an AI system more trustworthy to humans.

Topics of this lecture

- We briefly look at some methods:
  - Shapley: a method for feature attribution.
  - Probing: a method for post-hoc analysis.
  - Natural language explanation.

- We also look at how to evaluate the interpretability methods.

Basics for feature attribution

Recap: A feature-based model (i.e., linear regression).

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Assumption: **linear**, **feature-based** models are transparent.

Q: What is the *importance* of a feature, x_k?

A: We can attribute the prediction to x_k with β_k.

Shapley value ϕ_k

- Imagine n players are playing a game with y as the result.

- The Shapley value ϕ_k is the contribution / importance of x_k:
  - How much x_k can change y.

Lloyd Shapley won Nobel Memorial Prize in Economics in 2012

Meme: player importance

"We won without you.
We don't need you.
Leave."
-- Draymond Green

Shapley value ϕ_k

- Consider a set ("coalition") of players that do not contain x_k:

$$S \subseteq \{x_1 \dots x_n\} \backslash x_k$$

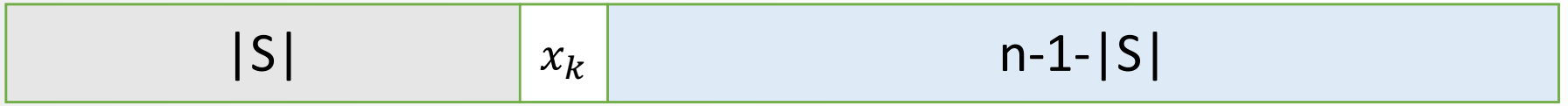
- How would the outcome y differ if these players play with x_k?

$$y_{S \cup x_k} - y_S$$

- The Shapley value ϕ_k is the *expectation* of such difference:

$$\phi_k = \mathbb{E}[y_{S \cup x_k} - y_S]$$

Shapley value ϕ_k



$$\phi_k = \mathbb{E}[y_{S \cup x_k} - y_S] = \sum_{S \in \{x_1 .. x_n\} \setminus x_k} p(S)[y_{S \cup x_k} - y_S]$$

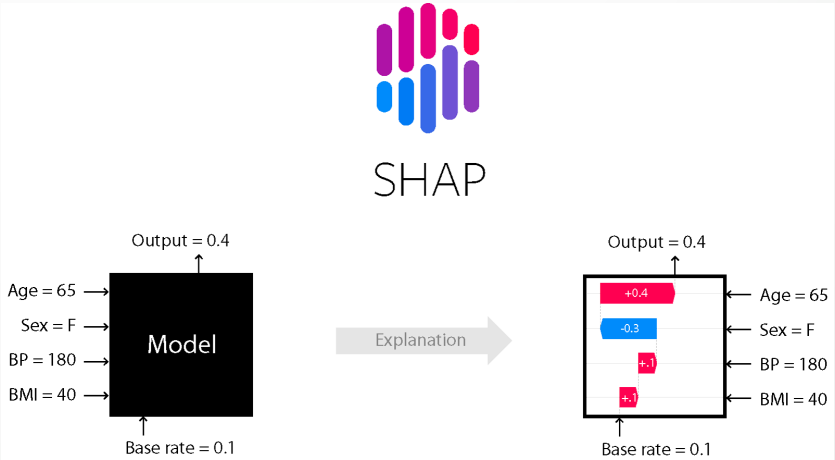
where:

$$p(S) = \frac{|S|!\,(n-1-|S|)!}{n!}$$

Aside: Computing Shapley value

- Exact computation is too expensive.
- Popular toolkits use Monte Carlo sampling.

$$\phi_k \approx \sum_{\text{Sample some } S} \frac{|S|!\,(n-1-|S|)!}{n!}\left[y_{S\cup x_k} - y_S\right]$$



Probing: post-hoc prediction

- Inspect the representations of the DNN models.
- Apply post-hoc prediction models (e.g., LogReg) to predict targets from the models' representations.
- Predictions can easily reach high performance

  ≈ the representations are informative.

Probing: post-hoc prediction

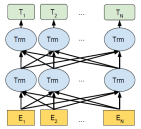
People refer to probing in two senses:

- A narrow sense: apply post-hoc ML models to the representations of DNNs.

- A broad sense: any approach that inspects the DNNs.

Probing are exams

Test

Participant: BERT 

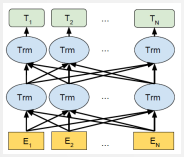
Test format: Linear probe

Result:

Ability observed on layer 4:

- Encode the syntax structure: 90% accuracy.
- Recognize the tense in a sentence: 95% accuracy.
- Detect the subject-verb agreement: 85% accuracy.
- …

Aside: Probing results are useful



probe

Tree Depth at layer 1: 0.70
Tense at layer 1: 0.85
SubjNum at layer 1: 0.50
…

fine-tune

QNLI accuracy: 0.95

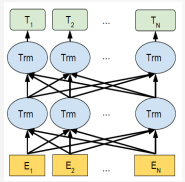
regress

This regression reaches RMSE much better than the regression from random features!

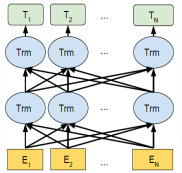
Multi-dimensional evaluations are helpful: Choose resuming checkpoints, hyperparameter tuning, etc.

Predicting fine-tuning performance with probing. Zhu et al., (2022) *EMNLP*

Aside: Probing reveals dataset effects



Multitask fine-tuning



Dataset effects

Test 1
- Syntax structure: 90%
- Recognize tense: 95%
- Subject-verb agreement: 85%

Test 2
- Syntax structure: 95%
- Recognize tense: 95%
- Subject-verb agreement: 80%

A state-vector framework for dataset effects. Sahak et al., *submitted*

Natural language explanations

- Question: An elephant can't be put into a fridge because it is too large. What is it?
  - (A) elephant
  - (B) fridge
- Answer: (A) elephant
- My explanation: An elephant is too large to be put into a fridge, so the pronoun "it" refers to the subject, "elephant".

Explanation provides common-sense knowledge.

Natural language explanations

- Question: Where do the winners of the 2018 Turing award work?
- Answer: University of Toronto, Université de Montréal, New York University.
- My explanation: The winners of the 2018 Turing award are Geoffrey Hinton, Yoshua Bengio, and Yann LeCun. They work at University of Toronto, Université de Montréal, New York University, respectively.

Explanation provides reasoning steps.

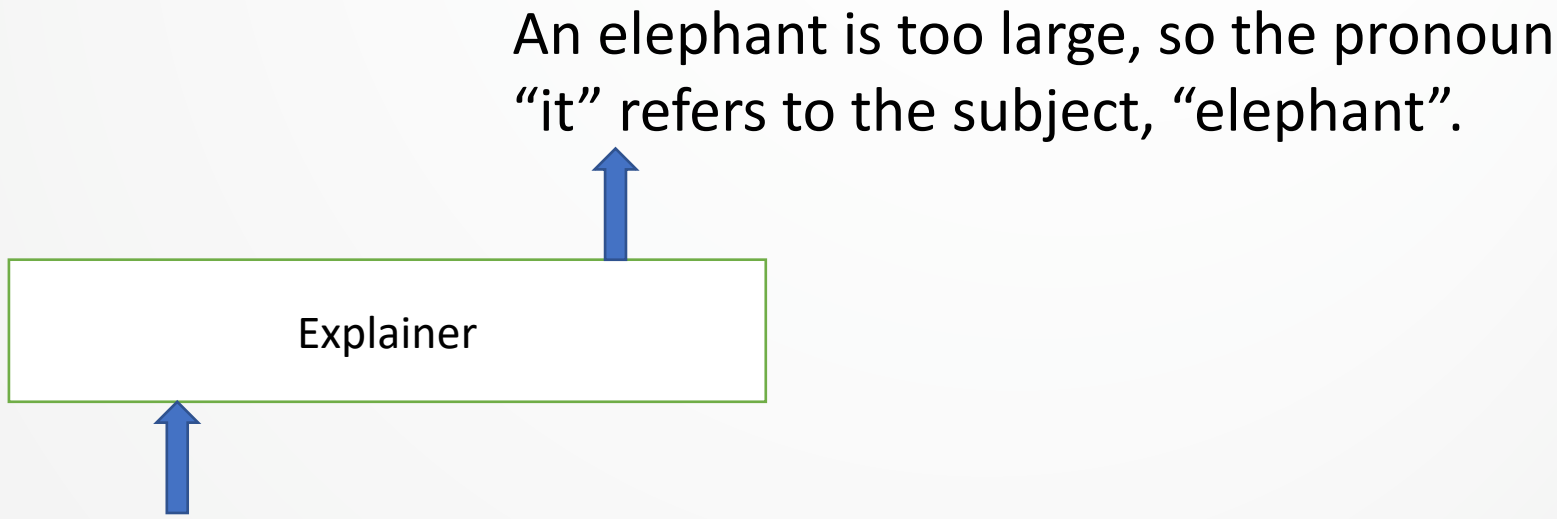
Natural language explanations

- Question: When I search for a bike, a helmet is returned among the list of results. Why?
- My explanation: A helmet makes bike rides safe.
  - Note: this explanation is high-level. It doesn't apply to everyone. I'll elaborate in a few slides.

Explanation provides post-hoc rationalization.

NLE as Seq2Seq generation

Here is an example and an expected explanation.



Needs some prompt engineering here

An elephant can't be put into a fridge because it is too large. It refers to the elephant because it is a metaphor for something that is too large to fit into a tight or small space.

Mode

Model

text-davinci-003

Temperature 0.7

ZI

An elephant can't be put into a fridge because it is too large. It refers to the elephant because

an elephant is too large to fit inside a standard-sized fridge. In general, it is not advisable to put large animals, such as elephants, inside fridges as they would not have enough room to move around and could potentially harm themselves or damage the fridge.

Aside: Situated explanations

Q: Searched for “bike”. Recommended a helmet.

A helmet makes bike rides safe.



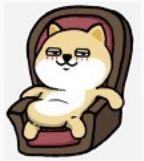

Situated Natural Language Explanations. Zhu et al., *Submitted*

Aside: Situated explanations

Q: Searched for “bike”. Recommended a helmet.

Because Ontario rules require bike riders below 18 to legally wear a helmet, and it is recommended that those above 18 should wear a helmet as well.



Situated Natural Language Explanations. Zhu et al., *Submitted*

Evaluating interpretability methods

- Model-centric evaluation
  - Do the explanations help improve model performances?
  - How much task-specific information is provided by the explanation?
  - How much do the performance change if we remove the highlighted features?

- Human-centric evaluation
  - Are humans more convinced about the decisions?
  - Does the explanation make the model more trustworthy?

Lecture review questions

By the end of this lecture, you should be able to:

- Describe Shapley value.
- Describe probing.
- Describe a method to generate natural language explanations.
- Describe how to evaluate interpretability methods.

Anonymous feedback form: https://forms.gle/W3i6AHaE4uRx2FAJA

UNIVERSITY OF TORONTO