HCXAI @ CHI 2021

Explaining the Road Not Taken

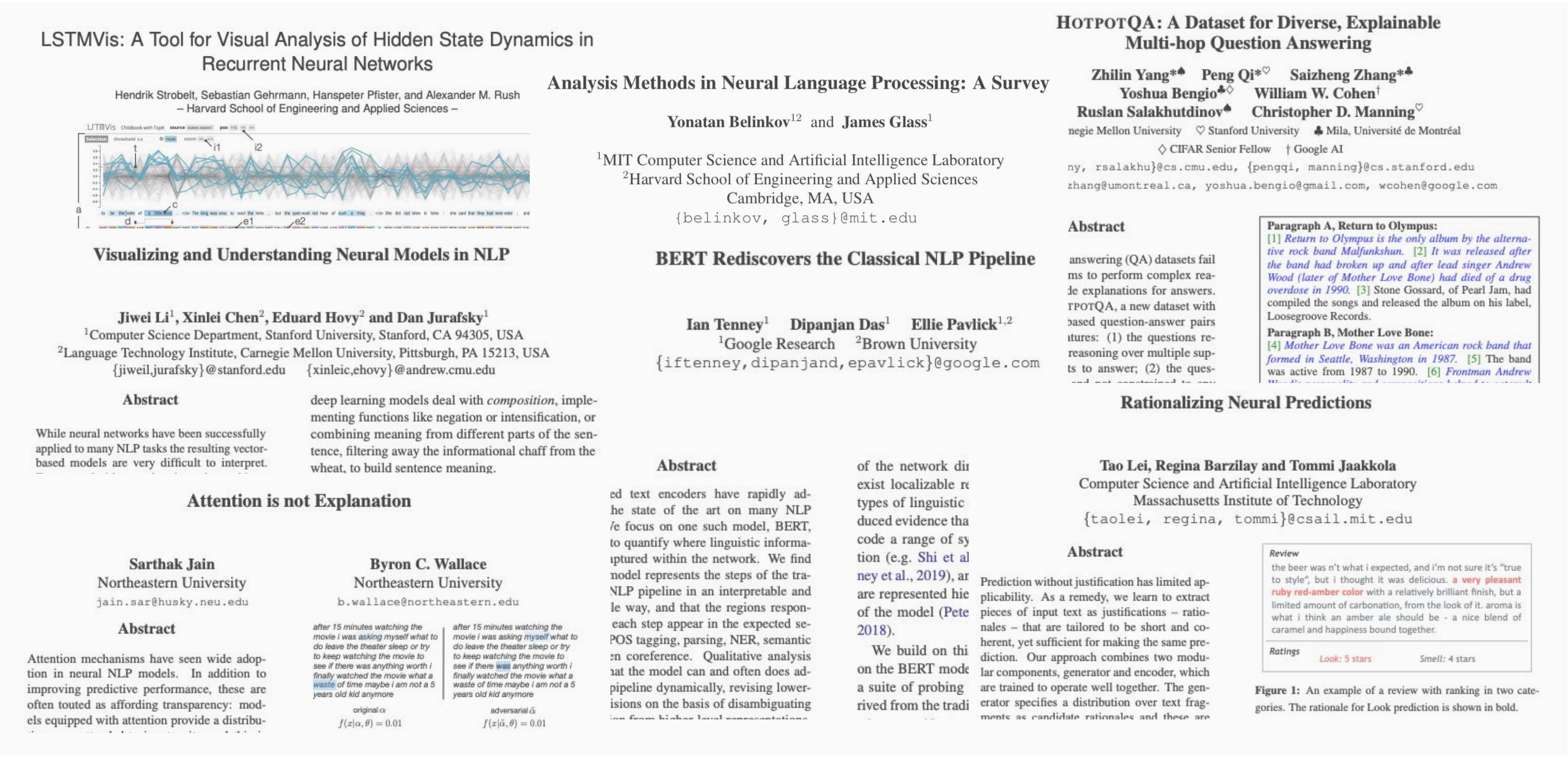
Hua Shen, Ting-Hao (Kenneth) Huang

The Pennsylvania State University

Crowd-AI Lab
crowdailab.net

PennState College of Information Sciences and Technology

NLP XAI Studies are Growing Rapidly



XAI Question Bank Shows Practical User Needs

Input

- **What kind of data does the system learn from?**
- What is the source of the data?
- How were the labels/ground-truth produced?
- * What is the sample size?
- * What data is the system NOT using?
- * What are the limitations/biases of the data?
- * How much data [like this] is the system trained on?

Output

- **What kind of output does the system give?**
- What does the system output mean?
- How can I best utilize the output of the system ?
- * What is the scope of the system's capability? Can it do…?
- * How is the output used for other system component(s) ?

Performance

- **How accurate/precise/reliable are the predictions?**
- How often does the system make mistakes?
- In what situations is the system likely to be correct/incorrect?
- * What are the limitations of the system?
- * What kind of mistakes is the system likely to make?
- * Is the system's performance good enough for…

How (global)

- **How does the system make predictions?**
- What features does the system consider?
  - * Is [feature X] used or not used for the predictions?
- What is the system's overall logic?
  - How does it weigh different features?
  - What rules does it use?
  - How does [feature X] impact its predictions?
  - * What are the top rules/features it uses?
- * What kind of algorithm is used?
  - * How are the parameters set?

Why

- **Why/how is this instance given this prediction?**
- What feature(s) of this instance leads to the system's prediction?
- Why are [instance A and B] given the same prediction?

Why not

- **Why/how is this instance NOT predicted…?**
- Why is this instance predicted P instead of Q?
- Why are [instance A and B] given different predictions?

What If

- **What would the system predict if this instance changes to…?**
- What would the system predict if this feature of the instance changes to…?
- What would the system predict for [a different instance]?

How to be that

- **How should this instance change to get a different prediction?**
- How should this feature change for this instance to get a different prediction?
- What kind of instance gets a different prediction?

How to still be this

- **What is the scope of change permitted to still get the same prediction?**
- What is the [highest/lowest/…] feature(s) one can have to still get the same prediction?
- What is the necessary feature(s) present or absent to guarantee this prediction?
- What kind of instance gets this prediction?

Others

- * How/what/why will the system change/adapt/improve/drift over time? (change)
- * How to improve the system? (change)
- * Why using or not using this feature/rule/data? (follow-up)
- * What does [ML terminology] mean? (terminological)
- * What are the results of other people using the system? (social)

(Liao, Q. V, et al, 2020)

How Well Can Existing NLP XAI Research Respond to these Questions that Users Care About?

We surveyed 200+ XAI Papers in NLP

ID	Title	Year	Venue	Paper URL
1	" Why should I trust you?" Explaining the predictions of any classifier	2016	KDD	https://arxiv.org/pdf/1602.04938.pdf
2	Visualizing and Understanding Neural Models in NLP	2016	NAACL	https://www.aclweb.org/anthology/N16-1082.pdf
3	Rationalizing Neural Predictions	2016	EMNLP	https://people.csail.mit.edu/taolei/papers/emnlp16_rationale.pdf
4	BERT Rediscovers the Classical NLP Pipeline	2019	ACL	https://www.aclweb.org/anthology/P19-1452.pdf
5	Attention is not Explanation	2019	NAACL	https://arxiv.org/pdf/1902.10186.pdf

Matching 200+ Papers with XAI Question Bank?

43 User Questions

Input

- **What kind of data does the system learn from?**
- What is the source of the data?
- How were the labels/ground-truth produced?
- * What is the sample size?
- * What data is the system NOT using?
- * What are the limitations/biases of the data?

⋮

Others

- * How/what/why will the system change/adapt/improve/drift over time? (change)
- * How to improve the system? (change)
- * Why using or not using this feature/rule/data? (follow-up)
- * What does [ML terminology] mean? (terminological)

?

218 NLP XAI Papers

ID	Title	Year	Venue	Paper URL
1	" Why should I trust you?" Explaining the predictions of any classifier	2016	KDD	https://arxiv.org/pdf/1602.04938.
2	Visualizing and Understanding Neural Models in NLP	2016	NAACL	https://www.aclweb.org/antholog
3	Rationalizing Neural Predictions	2016	EMNLP	https://people.csail.mit.edu/taole
4	BERT Rediscovers the Classical NLP Pipeline	2019	ACL	https://www.aclweb.org/antholog
5	Attention is not Explanation	2019	NAACL	https://arxiv.org/pdf/1902.10186.
⋮				
214	How much should you ask? On the question structure in QA systems	2018	BlackboxNLP	https://arxiv.org/pdf/1809.03734.
215	Interpretable Multi-dataset Evaluation for Named Entity Recognition	2020	EMNLP	https://arxiv.org/pdf/2011.06854.
216	A Survey of the State of Explainable AI for Natural Language Processin	2020	AACL-IJCNLP	https://arxiv.org/pdf/2010.00711.
217	Explaining Simple Natural Language Inference	2019	ACL	https://www.aclweb.org/antholog
218	Understanding Neural Abstractive Summarization Models via Uncertaint	2020	EMNLP	https://arxiv.org/pdf/2010.07882.

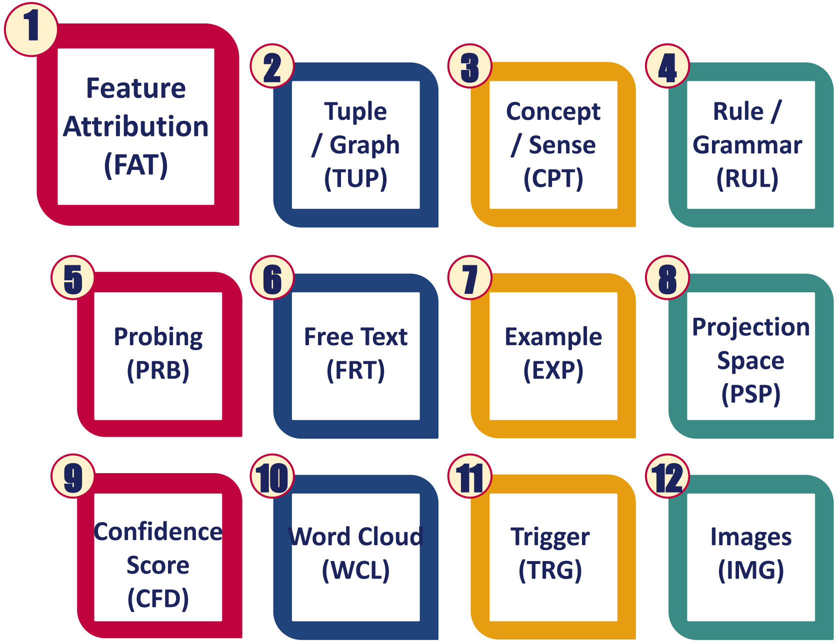
A Collection of XAI Forms

1. Feature Attribution (FAT)
2. Tuple / Graph (TUP)
3. Concept / Sense (CPT)
4. Rule / Grammar (RUL)
5. Probing (PRB)
6. Free Text (FRT)
7. Example (EXP)
8. Projection Space (PSP)
9. Confidence Score (CFD)
10. Word Cloud (WCL)
11. Trigger (TRG)
12. Images (IMG)

218 NLP XAI Papers

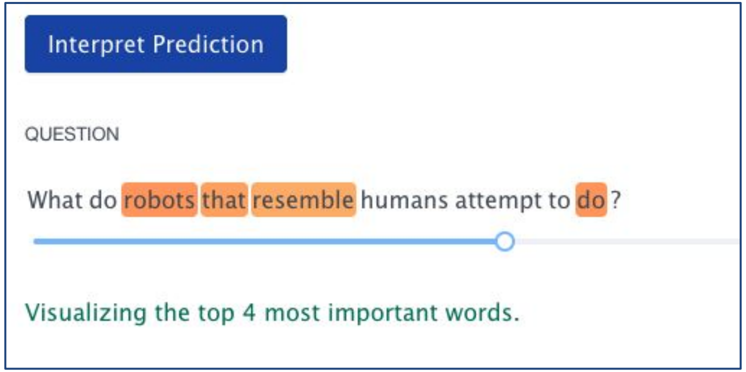
ID	Title	Year	Venue	Paper URL
1	" Why should I trust you?" Explaining the predictions of any classifier	2016	KDD	https://arxiv.org/pdf/1602.04938.
2	Visualizing and Understanding Neural Models in NLP	2016	NAACL	https://www.aclweb.org/antholog
3	Rationalizing Neural Predictions	2016	EMNLP	https://people.csail.mit.edu/taole
4	BERT Rediscovers the Classical NLP Pipeline	2019	ACL	https://www.aclweb.org/antholog
5	Attention is not Explanation	2019	NAACL	https://arxiv.org/pdf/1902.10186.
⋮				
214	How much should you ask? On the question structure in QA systems	2018	BlackboxNLP	https://arxiv.org/pdf/1809.03734.
215	Interpretable Multi-dataset Evaluation for Named Entity Recognition	2020	EMNLP	https://arxiv.org/pdf/2011.06854.
216	A Survey of the State of Explainable AI for Natural Language Processin	2020	AACL-IJCNLP	https://arxiv.org/pdf/2010.00711.
217	Explaining Simple Natural Language Inference	2019	ACL	https://www.aclweb.org/antholog
218	Understanding Neural Abstractive Summarization Models via Uncertaint	2020	EMNLP	https://arxiv.org/pdf/2010.07882.

A Collection of XAI Forms



1 - Feature Attribution (FAT)

- **Definition:** highlight the subsequence in input texts
- **Typical User Question:** "How can we attribute the AI systems' predictions to input features?"



(Wallace, Eric, et al, 2019)

Matching Each User Question with XAI Forms

XAI Question Bank

How (global) Q19 - How does the system make predictions?

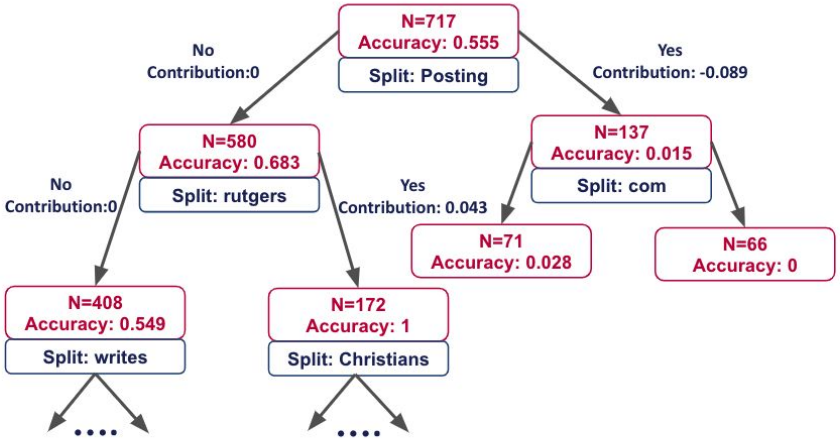
Matching Each User Question with XAI Forms

XAI Question Bank

How (global) Q19 - How does the system make predictions?

2

Tuple / Graph (TUP)



(Yang, C., et al, 2018)

Matching Each User Question with XAI Forms

XAI Question Bank

How (global) Q19 - How does the system make predictions?

4

Rule / Grammar (RUL)

if {“not”, “bad”} in input:
then **Positive**

if {“not”, “good”} in input:
then **Negative**

(Ribeiro, M. T., et al, 2018)

Matching **Each User Question** with **XAI Forms**

XAI Question Bank

How (global) **Q19 - How does the system make predictions?**

7

Example (EXP)

Positive Training Examples:

- This gem for gore lovers is extremely underrated. It's pure delight and fun! …..
- Project A II is a classic Jackie Chan movie with all the kung fu, crazy stunts and slapstick humor you expect…….

Negative Training Examples:

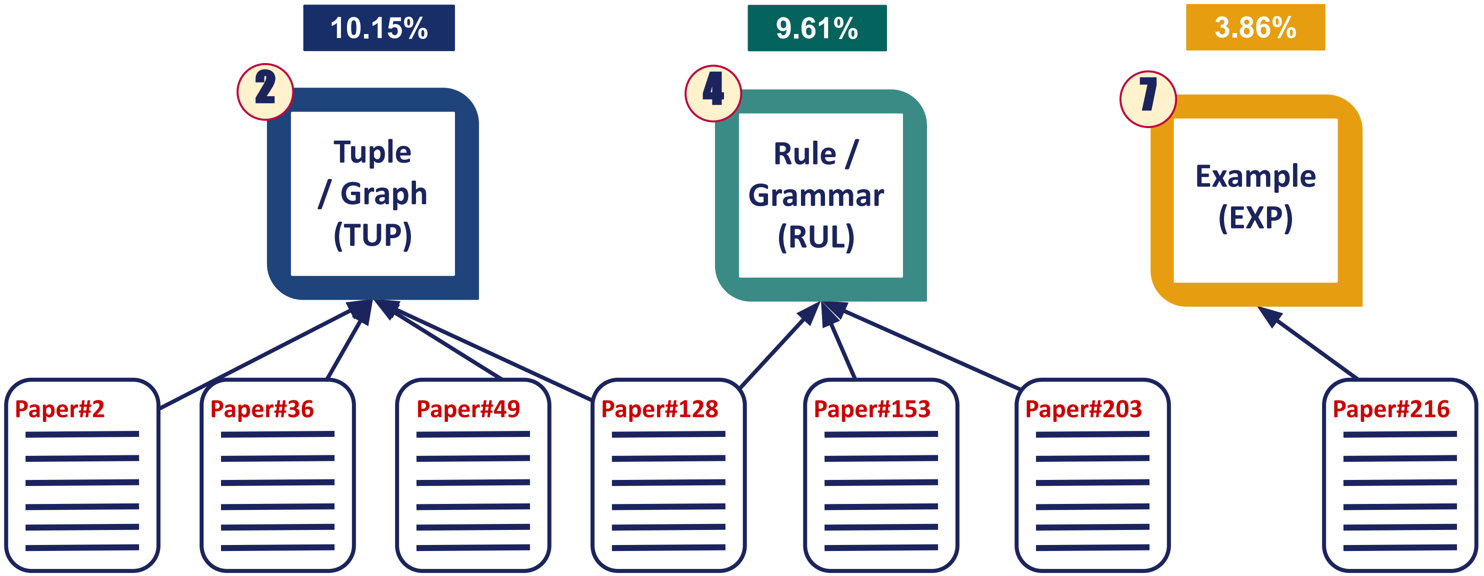
- Believe it or not, this was at one time the worst movie I had ever seen. …
- Great story and great lead actors (Quaid and Ryan) but the movie suffers from bad directing, bad screenplay and bad script…….

(Koh, P. W., & Liang, P, 2017)

Matching Each User Question with XAI Forms

XAI Question Bank

How (global) Q19 - How does the system make predictions?



Matching Each User Question with XAI Forms

XAI Question Bank

How (global) | Q19 - How does the system make predictions? → 23.62%

10.15% + 9.61% + 3.86% = 23.62%

2 Tuple / Graph (TUP)

4 Rule / Grammar (RUL)

7 Example (EXP)

Findings

Input/Data (0.55%)	1-What kind of data does the system learn from?	EXP	3.86%
	2-What is the source of the data?		★
	3-How were the labels/ground-truth produced?		★
	4-What is the sample size?		★
	5-What data is the system NOT using?		●
	6-What are the limitations/biases of the data?		●
	7-How much data [like this] is the system trained on?		★
Output (0.77%)	8-What kind of output does the system give?	EXP	3.86%
	9-What does the system output mean?		★
	10-How can I best utilize the output of the system?		●
	11-What is the scope of the system's capability?		●
	12-How's the output used for other systems modules?		●
Performance (2.03%)	13-How accurate/precise/reliable are the predictions?	CFD	1.18%
	14-How often does the system make mistakes?		★
	15-In what situations is the system to be incorrect?	CFD/EXP/TRG	5.97%
	16-What are the limitations of the system?		●
	17-What kind of mistake is the system likely to make?	EXP	5.05%
	18-Is the system's performance good enough for…?		●
How (Global) (30.31%)	19-How does the system make predictions?	TUP/RUL/EXP	23.63%
	20-What features does the system consider?	FAT	43.99%
	21-What is the system's overall logic?	RUL/FAT	53.60%
	22-What kind of algorithm is used?		★

Why / Why not (45.14%)	23-Why/how is this instance given this prediction?	RUL/TUP/FAT/FRT/EXP	74.70%
	24-What instance feature leads to the system's prediction?	FAT	43.99%
	25-Why are [instance A and B] given the same prediction?	RUL/TUP/FAT/FRT/EXP	74.70%
	26-Why/how is this instance NOT predicted?	TRG	0.93%
	27-Why is the instance predicted P instead of Q?	TRG	0.93%
	28-Why are [instance A and B] given different predictions?	TRG/RUL/TUP/FAT/FRT/EXP	75.62%
What if / How to be (15.54%)	29-What would the system predict if this instance changes to ..?	CFD/EXP/TRG	5.97%
	30-What would system predict if this instance feature changes to..?	CFD/FAT/TRG	46.10%
	31-What would the system predict for [a different instance]?	CFD/TRG	2.11%
	32-How should this instance change to get a different prediction?	TRG	0.93%
	33-How should instance feature change to get different prediction?	TRG	0.93%
	34-What kind of instance gets a different prediction?	TRG/EXP	4.79%
	35-What's the scope of change permitted to get the same prediction?	TRG	0.93%
	36-What's the highest feature can have to get the same prediction?	TRG/FAT	44.91%
	37-What is necessary feature present to guarantee this prediction?	TRG/FAT	44.91%
	38-What kind of instance gets this prediction?	EXP	3.86%
Others (11.49%)	39-How/what/why will the system change/improve/drift over time?		●
	40-How to improve the system?		●
	41-Why using or not using this feature/rule/data?	FAT/RUL/EXP	57.46%
	42-What does [ML terminology] mean?		★
	43-What are the results of other people using the system?		●

0.0% 100.0%

Findings

Category	Question	Method	Coverage
Input/Data (0.55%)	1-What kind of data does the system learn from?	EXP	3.86%
	2-What is the source of the data?		★
	3-How were the labels/ground-truth produced?		★
	4-What is the sample size?		★
	5-What data is the system NOT using?		●
	6-What are the limitations/biases of the data?		●
	7-How much data [like this] is the system trained on?		★
Output (0.77%)	8-What kind of output does the system give?	EXP	3.86%
	9-What does the system output mean?		★
	10-How can I best utilize the output of the system?		●
	11-What is the scope of the system's capability?		●
	12-How's the output used for other systems modules?		●
Performance (2.03%)	13-How accurate/precise/reliable are the predictions?	CFD	1.18%
	14-How often does the system make mistakes?		★
	15-In what situations is the system to be incorrect?	CFD/EXP/TRG	5.97%
	16-What are the limitations of the system?		●
	17-What kind of mistake is the system likely to make?	EXP	5.05%
	18-Is the system's performance good enough for…?		●
How (Global) (30.31%)	19-How does the system make predictions?	TUP/RUL/EXP	23.63%
	20-What features does the system consider?	FAT	43.99%
	21-What is the system's overall logic?	RUL/FAT	53.60%
	22-What kind of algorithm is used?		★
Why / Why not (45.14%)	23-Why/how is this instance given this prediction?	RUL/TUP/FAT/FRT/EXP	74.70%
	24-What instance feature leads to the system's prediction?	FAT	43.99%
	25-Why are [instance A and B] given the same prediction?	RUL/TUP/FAT/FRT/EXP	74.70%
	26-Why/how is this instance NOT predicted?	TRG	0.93%
	27-Why is the instance predicted P instead of Q?	TRG	0.93%
	28-Why are [instance A and B] given different predictions?	TRG/RUL/TUP/FAT/FRT/EXP	75.62%
What if / How to be (15.54%)	29-What would the system predict if this instance changes to ..?	CFD/EXP/TRG	5.97%
	30-What would system predict if this instance feature changes to..?	CFD/FAT/TRG	46.10%
	31-What would the system predict for [a different instance]?	CFD/TRG	2.11%
	32-How should this instance change to get a different prediction?	TRG	0.93%
	33-How should instance feature change to get different prediction?	TRG	0.93%
	34-What kind of instance gets a different prediction?	TRG/EXP	4.79%
	35-What's the scope of change permitted to get the same prediction?	TRG	0.93%
	36-What's the highest feature can have to get the same prediction?	TRG/FAT	44.91%
	37-What is necessary feature present to guarantee this prediction?	TRG/FAT	44.91%
	38-What kind of instance gets this prediction?	EXP	3.86%
Others (11.49%)	39-How/what/why will the system change/improve/drift over time?		●
	40-How to improve the system?		●
	41-Why using or not using this feature/rule/data?	FAT/RUL/EXP	57.46%
	42-What does [ML terminology] mean?		★
	43-What are the results of other people using the system?		●

0.0% — 100.0%

9 out of 43 questions: ***how AI systems CAN provide specific predictions***

Findings

Input/Data (0.55%)	1-What kind of data does the system learn from?	EXP	3.86%
	2-What is the source of the data?		★
	3-How were the labels/ground-truth produced?		★
	4-What is the sample size?		★
	5-What data is the system NOT using?		●
	6-What are the limitations/biases of the data?		●
	7-How much data [like this] is the system trained on?		★
Output (0.77%)	8-What kind of output does the system give?	EXP	3.86%
	9-What does the system output mean?		★
	10-How can I best utilize the output of the system?		●
	11-What is the scope of the system's capability?		●
	12-How's the output used for other systems modules?		●
Performance (2.03%)	13-How accurate/precise/reliable are the predictions?	CFD	1.18%
	14-How often does the system make mistakes?		★
	15-In what situations is the system to be incorrect?	CFD/EXP/TRG	5.97%
	16-What are the limitations of the system?		●
	17-What kind of mistake is the system likely to make?	EXP	5.05%
	18-Is the system's performance good enough for…?		●
How (Global) (30.31%)	19-How does the system make predictions?	TUP/RUL/EXP	23.63%
	20-What features does the system consider?	FAT	43.99%
	21-What is the system's overall logic?	RUL/FAT	53.60%
	22-What kind of algorithm is used?		★

Why / Why not (45.14%)	23-Why/how is this instance given this prediction?	RUL/TUP/FAT/FRT/EXP	74.70%
	24-What instance feature leads to the system's prediction?	FAT	43.99%
	25-Why are [instance A and B] given the same prediction?	RUL/TUP/FAT/FRT/EXP	74.70%
	26-Why/how is this instance NOT predicted?	TRG	0.93%
	27-Why is the instance predicted P instead of Q?	TRG	0.93%
	28-Why are [instance A and B] given different predictions?	TRG/RUL/TUP/FAT/FRT/EXP	75.62%
What if / How to be (15.54%)	29-What would the system predict if this instance changes to ..?	CFD/EXP/TRG	5.97%
	30-What would system predict if this instance feature changes to..?	CFD/FAT/TRG	46.10%
	31-What would the system predict for [a different instance]?	CFD/TRG	2.11%
	32-How should this instance change to get a different prediction?	TRG	0.93%
	33-How should instance feature change to get different prediction?	TRG	0.93%
	34-What kind of instance gets a different prediction?	TRG/EXP	4.79%
	35-What's the scope of change permitted to get the same prediction?	TRG	0.93%
	36-What's the highest feature can have to get the same prediction?	TRG/FAT	44.91%
	37-What is necessary feature present to guarantee this prediction?	TRG/FAT	44.91%
	38-What kind of instance gets this prediction?	EXP	3.86%
Others (11.49%)	39-How/what/why will the system change/improve/drift over time?		●
	40-How to improve the system?		●
	41-Why using or not using this feature/rule/data?	FAT/RUL/EXP	57.46%
	42-What does [ML terminology] mean?		★
	43-What are the results of other people using the system?		●

0.0% — 100.0%

➡ **16 out of 43 questions: *what AI systems CANNOT achieve***

Explaining the Road Not Taken

Users are **interested** in explanations for the ***road not taken*** -- namely, why AI chose current prediction **instead of** a **legitimate counterpart**

Website: https://human-centered-exnlp.github.io/

Open 200+ NLP Explanation Form Annotations

ID	Title	Year	Venue (Abbreviation List)	Feature Attribution (FAT)	Probing (PRB)	Tuple/Graph (TUP)	Projection Space (PSP)	Rule/Grammar (RUL)	Free Text (FRT)	Concept/Sense (CPT)	Example (EXP)	Trigger (TRG)	Word Cloud (WCL)	Images (IMG)	Confidence Score (CFD)
1	" Why should I trust you?" Explaining the predictions of any classifier	2016	KDD	Quote	-	-	-	-	-	-	-	-	-	-	-
2	A causal framework for explaining the predictions of black-box sequence-to-sequence models	2017	EMNLP	Quote	-	Quote	-	-	-	-	-	-	-	-	-
3	A Diagnostic Study of Explainability Techniques for Text Classification	2020	EMNLP	Quote	-	-	-	-	-	-	-	-	-	-	-
4	A Meaning-based English Math Word Problem Solver with Understanding, Reasoning and Explanation	2016	COLING	-	-	-	-	Quote	Quote	-	-	-	-	-	-
5	A primer in bertology: What we know about how bert works	2020	TACL	Quote	-	Quote	-	-	-	-	-	-	-	-	-
6	A Shared Attention Mechanism for Interpretation of Neural Automatic Post-Editing Systems	2018	ACL	Quote	-	-	-	-	-	-	-	-	-	-	-
7	A structural probe for finding syntax in word representations	2019	NAACL	-	-	Quote	-	-	-	-	-	-	-	-	-
8	A Survey of the State of Explainable AI for Natural Language Processing	2020	AACL-IJCNLP	Quote	-	-	-	Quote	Quote	-	-	-	-	-	-
9	Allennlp interpret: A framework for explaining predictions of nlp models	2019	EMNLP	Quote	-	-	-	Quote	-	-	-	-	-	-	-
10	An Information Bottleneck Approach for Controlling Conciseness in Rationale Extraction	2020	EMNLP	Quote	-	-	-	-	-	-	-	-	-	-	-

Thank you!

Hua Shen

huashen218@psu.edu

@SarahHShen1

Ting-Hao (Kenneth) Huang

txh710@psu.edu

@windx0303