

# Explaining the Road Not Taken

**Hua Shen**, Ting-Hao (Kenneth) Huang

The Pennsylvania State University



**Crowd-AI Lab**  
crowdailab.net

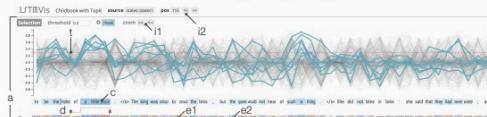


**PennState**  
College of Information  
Sciences and Technology

# NLP XAI Studies are Growing Rapidly

## LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks

Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M. Rush  
—Harvard School of Engineering and Applied Sciences—



### Visualizing and Understanding Neural Models in NLP

Jiwei Li<sup>1</sup>, Xinlei Chen<sup>2</sup>, Eduard Hovy<sup>2</sup> and Dan Jurafsky<sup>1</sup>

<sup>1</sup>Computer Science Department, Stanford University, Stanford, CA 94305, USA

<sup>2</sup>Language Technology Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA

{jiweili, jurafsky}@stanford.edu {xinleic, ehovy}@andrew.cmu.edu

### Abstract

While neural networks have been successfully applied to many NLP tasks the resulting vector-based models are very difficult to interpret.

deep learning models deal with *composition*, implementing functions like negation or intensification, or combining meaning from different parts of the sentence, filtering away the informational chaff from the wheat, to build sentence meaning.

### Attention is not Explanation

Sarthak Jain

Northeastern University

jain.sar@husky.neu.edu

### Abstract

Attention mechanisms have seen wide adoption in neural NLP models. In addition to improving predictive performance, these are often touted as affording transparency: models equipped with attention provide a distribu-

Byron C. Wallace

Northeastern University

b.wallace@northeastern.edu

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

original  $\hat{\alpha}$   
 $f(x|\alpha, \theta) = 0.01$

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

adversarial  $\hat{\alpha}$   
 $f(x|\hat{\alpha}, \theta) = 0.01$

## Analysis Methods in Neural Language Processing: A Survey

Yonatan Belinkov<sup>1,2</sup> and James Glass<sup>1</sup>

<sup>1</sup>MIT Computer Science and Artificial Intelligence Laboratory

<sup>2</sup>Harvard School of Engineering and Applied Sciences

Cambridge, MA, USA

{belinkov, glass}@mit.edu

## BERT Rediscovered the Classical NLP Pipeline

Ian Tenney<sup>1</sup> Dipanjan Das<sup>1</sup> Ellie Pavlick<sup>1,2</sup>

<sup>1</sup>Google Research <sup>2</sup>Brown University

{iftenney, dipanjan, epavlick}@google.com

### Abstract

ed text encoders have rapidly achieved state-of-the-art on many NLP tasks. To quantify where linguistic information is captured within the network, we find model representations of the steps of the NLP pipeline in an interpretable and learnable way, and that the regions responsible for each step appear in the expected sequence of POS tagging, parsing, NER, semantic role labeling, and coreference. Qualitative analysis of the model can and often does adapt the pipeline dynamically, revising lower-level decisions on the basis of disambiguating information from higher-level representations.

of the network do exist localizable representations of linguistic information. We find evidence that the model codes a range of syntactic information (e.g. Shi et al., 2019), and that the regions responsible for each step appear in the expected sequence of the model (Pete et al., 2018).

We build on this with the BERT model as a suite of probing tasks derived from the tradi-

## HOTPOTQA: A Dataset for Diverse, Explainable Multi-hop Question Answering

Zhilin Yang<sup>\*†</sup> Peng Qi<sup>\*‡</sup> Saizheng Zhang<sup>\*‡</sup>

Yoshua Bengio<sup>\*§</sup> William W. Cohen<sup>†</sup>

Ruslan Salakhutdinov<sup>†</sup> Christopher D. Manning<sup>‡</sup>

negie Mellon University <sup>‡</sup>Stanford University <sup>§</sup>Mila, Université de Montréal

<sup>†</sup>CIFAR Senior Fellow <sup>‡</sup>Google AI

ny, rsalakh@cs.cmu.edu, {pengqi, manning}@cs.stanford.edu

zhang@umontreal.ca, yoshua.bengio@gmail.com, wcohen@google.com

### Abstract

answering (QA) datasets fail to perform complex reasoning for answers. HOTPOTQA, a new dataset with complex question-answer pairs, features: (1) the questions requiring reasoning over multiple supports to answer; (2) the ques-

### Paragraph A. Return to Olympus:

[1] Return to Olympus is the only album by the alternative rock band Malfunkshun. [2] It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990. [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

### Paragraph B. Mother Love Bone:

[4] Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987. [5] The band was active from 1987 to 1990. [6] Frontman Andrew

## Rationalizing Neural Predictions

Tao Lei, Regina Barzilay and Tommi Jaakkola

Computer Science and Artificial Intelligence Laboratory

Massachusetts Institute of Technology

{taolei, regina, tommi}@csail.mit.edu

### Abstract

Prediction without justification has limited applicability. As a remedy, we learn to extract pieces of input text as justifications – rationales – that are tailored to be short and coherent, yet sufficient for making the same prediction. Our approach combines two modular components, generator and encoder, which are trained to operate well together. The generator specifies a distribution over text fragments as candidate rationales and these are

### Review

the beer was n't what i expected, and i'm not sure it's "true to style", but i thought it was delicious. a very pleasant ruby red-amber color with a relatively brilliant finish, but a limited amount of carbonation, from the look of it. aroma is what i think an amber ale should be - a nice blend of caramel and happiness bound together.

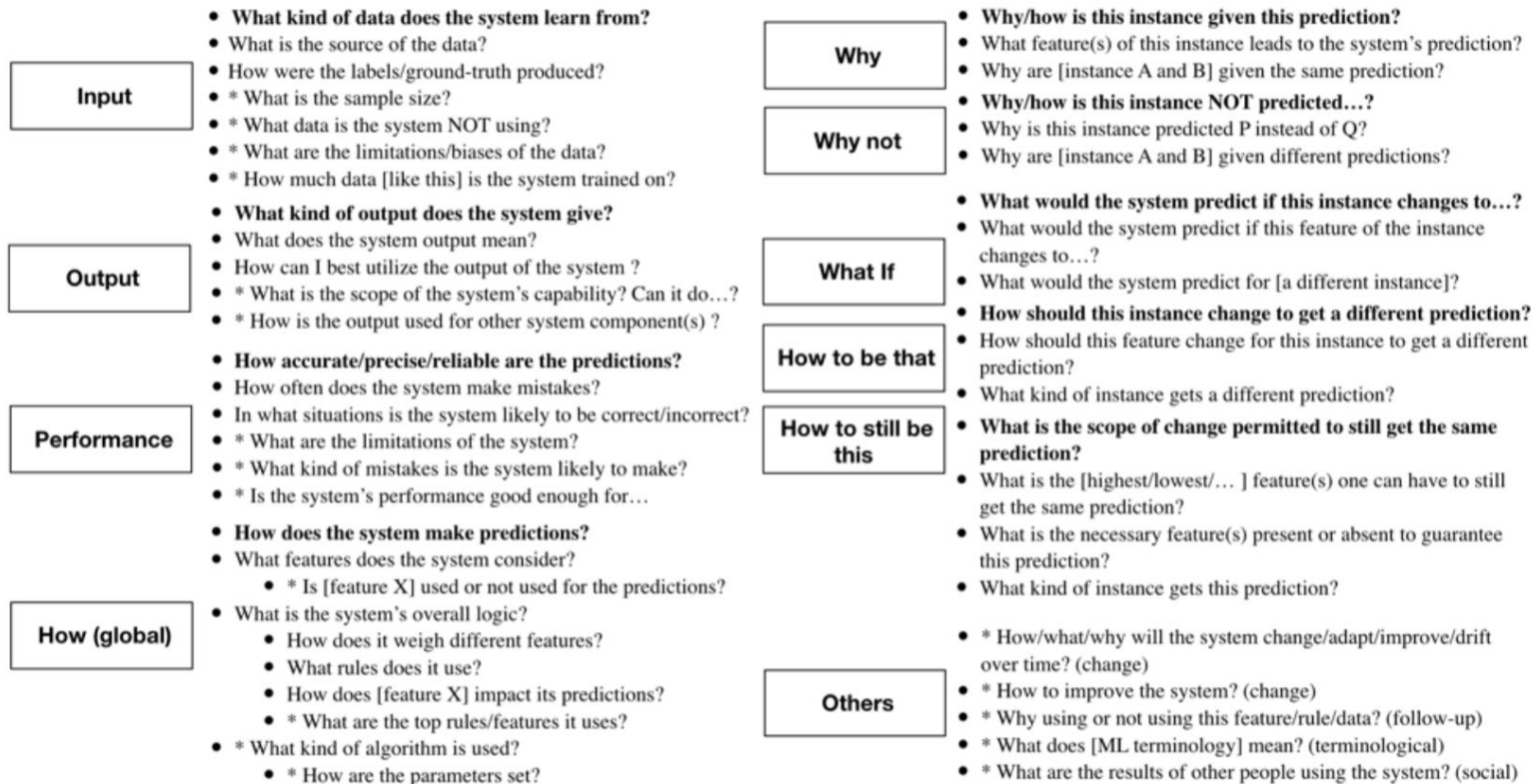
### Ratings

Look: 5 stars

Smell: 4 stars

Figure 1: An example of a review with ranking in two categories. The rationale for Look prediction is shown in bold.

# XAI Question Bank Shows Practical User Needs



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



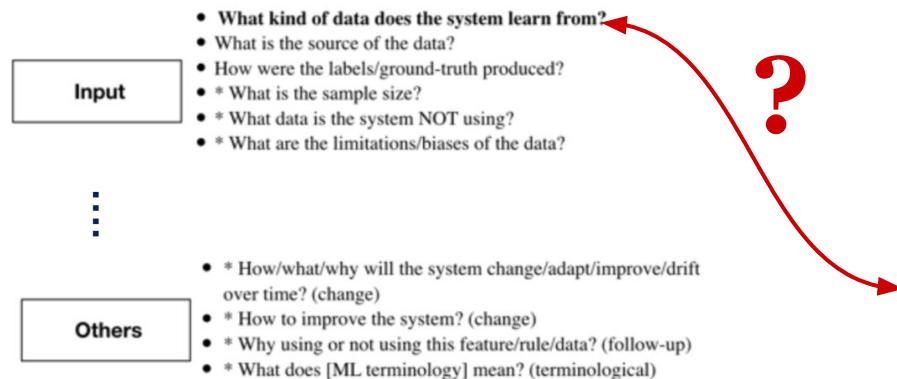
# We surveyed 200+ XAI Papers in NLP

ID	Title	Year	Venue	Paper URL	ID	Title	Year	Venue	Paper URL
1	"Why should I trust you?" Explaining the predictions of any classifier	2016	KDD	<a href="https://arxiv.org/pdf/1602.04938.pdf">https://arxiv.org/pdf/1602.04938.pdf</a>	17	Multi-modal language analysis in the wild: On-mono dataset and multi-modal	2018	ACL	<a href="https://arxiv.org/pdf/1802.07282.pdf">https://arxiv.org/pdf/1802.07282.pdf</a>
2	Visualizing and Understanding Neural Models in NLP	2016	NAACL	<a href="https://www.aclweb.org/anthology/N16-1082.pdf">https://www.aclweb.org/anthology/N16-1082.pdf</a>	18	MultiModal Reasoning: Improving Local and Global Interpretability of Interpretable	2018	EMNLP	<a href="https://arxiv.org/pdf/2020.14198.pdf">https://arxiv.org/pdf/2020.14198.pdf</a>
3	Rationalizing Neural Predictions	2016	EMNLP	<a href="https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf">https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf</a>	19	Natural Language Reasoning with Full-Scale Visual Reasoning: From Pixels to	2018	EMNLP	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
4	BERT Rediscovers the Classical NLP Pipeline	2019	ACL	<a href="https://www.aclweb.org/anthology/P19-1452.pdf">https://www.aclweb.org/anthology/P19-1452.pdf</a>	20	Neural vector contextualization for word vector space interpretation	2019	NAACL	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
5	Attention is not Explanation	2019	NAACL	<a href="https://arxiv.org/pdf/1902.10186.pdf">https://arxiv.org/pdf/1902.10186.pdf</a>	21	No Explanation without Accountability: An Empirical Study of Explanations	2020	CHI	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
6	Why should I trust you? Explaining the predictions of any classifier	2016	KDD	<a href="https://arxiv.org/pdf/1602.04938.pdf">https://arxiv.org/pdf/1602.04938.pdf</a>	22	Open Sesame: Getting Inside BERT's Linguistic Knowledge	2019	BlackBoxNLP	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
7	Visualizing and Understanding Neural Models in NLP	2016	NAACL	<a href="https://www.aclweb.org/anthology/N16-1082.pdf">https://www.aclweb.org/anthology/N16-1082.pdf</a>	23	Open Sesame: Getting Inside BERT's Linguistic Knowledge	2019	ACL	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
8	Rationalizing Neural Predictions	2016	EMNLP	<a href="https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf">https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf</a>	24	Open Sesame: Getting Inside BERT's Linguistic Knowledge	2019	ACL	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
9	BERT Rediscovers the Classical NLP Pipeline	2019	ACL	<a href="https://www.aclweb.org/anthology/P19-1452.pdf">https://www.aclweb.org/anthology/P19-1452.pdf</a>	25	Pathologies of Neural Models Make Interpretations Difficult	2018	EMNLP	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
10	Attention is not Explanation	2019	NAACL	<a href="https://arxiv.org/pdf/1902.10186.pdf">https://arxiv.org/pdf/1902.10186.pdf</a>	26	Paraphrase Masking: Parameter-free Probing for Analyzing and Interpreting BERT	2018	ACL	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
11	Why should I trust you? Explaining the predictions of any classifier	2016	KDD	<a href="https://arxiv.org/pdf/1602.04938.pdf">https://arxiv.org/pdf/1602.04938.pdf</a>	27	Predicting and Interpreting Embeddings for out of vocabulary words in down	2018	BlackBoxNLP	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
12	Visualizing and Understanding Neural Models in NLP	2016	NAACL	<a href="https://www.aclweb.org/anthology/N16-1082.pdf">https://www.aclweb.org/anthology/N16-1082.pdf</a>	28	Principles of Explanatory Debating to Personalize Interactive Machine Learning	2019	CHI	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
13	Rationalizing Neural Predictions	2016	EMNLP	<a href="https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf">https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf</a>	29	Probing Emergent Semantics in Predictive Agents via Question Answering	2019	Anv	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
14	BERT Rediscovers the Classical NLP Pipeline	2019	ACL	<a href="https://www.aclweb.org/anthology/P19-1452.pdf">https://www.aclweb.org/anthology/P19-1452.pdf</a>	30	Probing for semantic evidence of composition by means of simple classification	2016	EMNLP	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
15	Attention is not Explanation	2019	NAACL	<a href="https://arxiv.org/pdf/1902.10186.pdf">https://arxiv.org/pdf/1902.10186.pdf</a>	31	Probing Neural Design Models for Conversational Understanding	2018	ACL	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
16	Why should I trust you? Explaining the predictions of any classifier	2016	KDD	<a href="https://arxiv.org/pdf/1602.04938.pdf">https://arxiv.org/pdf/1602.04938.pdf</a>	32	Principles of Explanatory Debating to Personalize Interactive Machine Learning	2019	CHI	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
17	Visualizing and Understanding Neural Models in NLP	2016	NAACL	<a href="https://www.aclweb.org/anthology/N16-1082.pdf">https://www.aclweb.org/anthology/N16-1082.pdf</a>	33	PROVER: A Framework for Generation for Interpretable Reasoning over Just	2020	EMNLP	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
18	Rationalizing Neural Predictions	2016	EMNLP	<a href="https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf">https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf</a>	34	Quick and (not so) Accurate: Supervised Selection of Justification Sentences	2020	EMNLP	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
19	BERT Rediscovers the Classical NLP Pipeline	2019	ACL	<a href="https://www.aclweb.org/anthology/P19-1452.pdf">https://www.aclweb.org/anthology/P19-1452.pdf</a>	35	Principles of Explanatory Debating to Personalize Interactive Machine Learning	2019	CHI	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
20	Attention is not Explanation	2019	NAACL	<a href="https://arxiv.org/pdf/1902.10186.pdf">https://arxiv.org/pdf/1902.10186.pdf</a>	36	Rationalizing Neural Predictions	2016	EMNLP	<a href="https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf">https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf</a>
21	Why should I trust you? Explaining the predictions of any classifier	2016	KDD	<a href="https://arxiv.org/pdf/1602.04938.pdf">https://arxiv.org/pdf/1602.04938.pdf</a>	37	Rethinking Cooperative Rationalization: Interprocedural Reasoning and Compi	2019	EMNLP	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
22	Visualizing and Understanding Neural Models in NLP	2016	NAACL	<a href="https://www.aclweb.org/anthology/N16-1082.pdf">https://www.aclweb.org/anthology/N16-1082.pdf</a>	38	Self-Driven word alignment interpretation for neural machine translation	2019	ACL	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
23	Rationalizing Neural Predictions	2016	EMNLP	<a href="https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf">https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf</a>	39	Self-Attention Module Networks for Interpretable Multi-Hop Reasoning	2019	EMNLP	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
24	BERT Rediscovers the Classical NLP Pipeline	2019	ACL	<a href="https://www.aclweb.org/anthology/P19-1452.pdf">https://www.aclweb.org/anthology/P19-1452.pdf</a>	40	Self-Critical Reasoning for Robust Visual Question Answering	2019	EMNLP	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
25	Attention is not Explanation	2019	NAACL	<a href="https://arxiv.org/pdf/1902.10186.pdf">https://arxiv.org/pdf/1902.10186.pdf</a>	41	Self-Explaining Structures Improve NLP Models	2020	Anv	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
26	Why should I trust you? Explaining the predictions of any classifier	2016	KDD	<a href="https://arxiv.org/pdf/1602.04938.pdf">https://arxiv.org/pdf/1602.04938.pdf</a>	42	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
27	Visualizing and Understanding Neural Models in NLP	2016	NAACL	<a href="https://www.aclweb.org/anthology/N16-1082.pdf">https://www.aclweb.org/anthology/N16-1082.pdf</a>	43	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
28	Rationalizing Neural Predictions	2016	EMNLP	<a href="https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf">https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf</a>	44	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
29	BERT Rediscovers the Classical NLP Pipeline	2019	ACL	<a href="https://www.aclweb.org/anthology/P19-1452.pdf">https://www.aclweb.org/anthology/P19-1452.pdf</a>	45	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
30	Attention is not Explanation	2019	NAACL	<a href="https://arxiv.org/pdf/1902.10186.pdf">https://arxiv.org/pdf/1902.10186.pdf</a>	46	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
31	Why should I trust you? Explaining the predictions of any classifier	2016	KDD	<a href="https://arxiv.org/pdf/1602.04938.pdf">https://arxiv.org/pdf/1602.04938.pdf</a>	47	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
32	Visualizing and Understanding Neural Models in NLP	2016	NAACL	<a href="https://www.aclweb.org/anthology/N16-1082.pdf">https://www.aclweb.org/anthology/N16-1082.pdf</a>	48	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
33	Rationalizing Neural Predictions	2016	EMNLP	<a href="https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf">https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf</a>	49	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
34	BERT Rediscovers the Classical NLP Pipeline	2019	ACL	<a href="https://www.aclweb.org/anthology/P19-1452.pdf">https://www.aclweb.org/anthology/P19-1452.pdf</a>	50	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
35	Attention is not Explanation	2019	NAACL	<a href="https://arxiv.org/pdf/1902.10186.pdf">https://arxiv.org/pdf/1902.10186.pdf</a>	51	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
36	Why should I trust you? Explaining the predictions of any classifier	2016	KDD	<a href="https://arxiv.org/pdf/1602.04938.pdf">https://arxiv.org/pdf/1602.04938.pdf</a>	52	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
37	Visualizing and Understanding Neural Models in NLP	2016	NAACL	<a href="https://www.aclweb.org/anthology/N16-1082.pdf">https://www.aclweb.org/anthology/N16-1082.pdf</a>	53	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
38	Rationalizing Neural Predictions	2016	EMNLP	<a href="https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf">https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf</a>	54	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
39	BERT Rediscovers the Classical NLP Pipeline	2019	ACL	<a href="https://www.aclweb.org/anthology/P19-1452.pdf">https://www.aclweb.org/anthology/P19-1452.pdf</a>	55	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
40	Attention is not Explanation	2019	NAACL	<a href="https://arxiv.org/pdf/1902.10186.pdf">https://arxiv.org/pdf/1902.10186.pdf</a>	56	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
41	Why should I trust you? Explaining the predictions of any classifier	2016	KDD	<a href="https://arxiv.org/pdf/1602.04938.pdf">https://arxiv.org/pdf/1602.04938.pdf</a>	57	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
42	Visualizing and Understanding Neural Models in NLP	2016	NAACL	<a href="https://www.aclweb.org/anthology/N16-1082.pdf">https://www.aclweb.org/anthology/N16-1082.pdf</a>	58	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
43	Rationalizing Neural Predictions	2016	EMNLP	<a href="https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf">https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf</a>	59	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
44	BERT Rediscovers the Classical NLP Pipeline	2019	ACL	<a href="https://www.aclweb.org/anthology/P19-1452.pdf">https://www.aclweb.org/anthology/P19-1452.pdf</a>	60	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
45	Attention is not Explanation	2019	NAACL	<a href="https://arxiv.org/pdf/1902.10186.pdf">https://arxiv.org/pdf/1902.10186.pdf</a>	61	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
46	Why should I trust you? Explaining the predictions of any classifier	2016	KDD	<a href="https://arxiv.org/pdf/1602.04938.pdf">https://arxiv.org/pdf/1602.04938.pdf</a>	62	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
47	Visualizing and Understanding Neural Models in NLP	2016	NAACL	<a href="https://www.aclweb.org/anthology/N16-1082.pdf">https://www.aclweb.org/anthology/N16-1082.pdf</a>	63	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
48	Rationalizing Neural Predictions	2016	EMNLP	<a href="https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf">https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf</a>	64	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
49	BERT Rediscovers the Classical NLP Pipeline	2019	ACL	<a href="https://www.aclweb.org/anthology/P19-1452.pdf">https://www.aclweb.org/anthology/P19-1452.pdf</a>	65	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
50	Attention is not Explanation	2019	NAACL	<a href="https://arxiv.org/pdf/1902.10186.pdf">https://arxiv.org/pdf/1902.10186.pdf</a>	66	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
51	Why should I trust you? Explaining the predictions of any classifier	2016	KDD	<a href="https://arxiv.org/pdf/1602.04938.pdf">https://arxiv.org/pdf/1602.04938.pdf</a>	67	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
52	Visualizing and Understanding Neural Models in NLP	2016	NAACL	<a href="https://www.aclweb.org/anthology/N16-1082.pdf">https://www.aclweb.org/anthology/N16-1082.pdf</a>	68	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
53	Rationalizing Neural Predictions	2016	EMNLP	<a href="https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf">https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf</a>	69	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
54	BERT Rediscovers the Classical NLP Pipeline	2019	ACL	<a href="https://www.aclweb.org/anthology/P19-1452.pdf">https://www.aclweb.org/anthology/P19-1452.pdf</a>	70	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
55	Attention is not Explanation	2019	NAACL	<a href="https://arxiv.org/pdf/1902.10186.pdf">https://arxiv.org/pdf/1902.10186.pdf</a>	71	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
56	Why should I trust you? Explaining the predictions of any classifier	2016	KDD	<a href="https://arxiv.org/pdf/1602.04938.pdf">https://arxiv.org/pdf/1602.04938.pdf</a>	72	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
57	Visualizing and Understanding Neural Models in NLP	2016	NAACL	<a href="https://www.aclweb.org/anthology/N16-1082.pdf">https://www.aclweb.org/anthology/N16-1082.pdf</a>	73	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
58	Rationalizing Neural Predictions	2016	EMNLP	<a href="https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf">https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf</a>	74	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
59	BERT Rediscovers the Classical NLP Pipeline	2019	ACL	<a href="https://www.aclweb.org/anthology/P19-1452.pdf">https://www.aclweb.org/anthology/P19-1452.pdf</a>	75	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
60	Attention is not Explanation	2019	NAACL	<a href="https://arxiv.org/pdf/1902.10186.pdf">https://arxiv.org/pdf/1902.10186.pdf</a>	76	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
61	Why should I trust you? Explaining the predictions of any classifier	2016	KDD	<a href="https://arxiv.org/pdf/1602.04938.pdf">https://arxiv.org/pdf/1602.04938.pdf</a>	77	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
62	Visualizing and Understanding Neural Models in NLP	2016	NAACL	<a href="https://www.aclweb.org/anthology/N16-1082.pdf">https://www.aclweb.org/anthology/N16-1082.pdf</a>	78	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
63	Rationalizing Neural Predictions	2016	EMNLP	<a href="https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf">https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf</a>	79	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
64	BERT Rediscovers the Classical NLP Pipeline	2019	ACL	<a href="https://www.aclweb.org/anthology/P19-1452.pdf">https://www.aclweb.org/anthology/P19-1452.pdf</a>	80	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
65	Attention is not Explanation	2019	NAACL	<a href="https://arxiv.org/pdf/1902.10186.pdf">https://arxiv.org/pdf/1902.10186.pdf</a>	81	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
66	Why should I trust you? Explaining the predictions of any classifier	2016	KDD	<a href="https://arxiv.org/pdf/1602.04938.pdf">https://arxiv.org/pdf/1602.04938.pdf</a>	82	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
67	Visualizing and Understanding Neural Models in NLP	2016	NAACL	<a href="https://www.aclweb.org/anthology/N16-1082.pdf">https://www.aclweb.org/anthology/N16-1082.pdf</a>	83	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
68	Rationalizing Neural Predictions	2016	EMNLP	<a href="https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf">https://people.csail.mit.edu/taoie/papers/emnlp16_rationale.pdf</a>	84	Explainable AI: A survey of explainable artificial intelligence	2019	TYCV	<a href="https://arxiv.org/pdf/2020.07286.pdf">https://arxiv.org/pdf/2020.07286.pdf</a>
69	BERT Rediscovers the Classical NLP Pipeline	2019	ACL	<a href="https://www.aclweb.org/anthology/P19-1452.pdf">https://www.aclweb.org/anthology/P19-1452.pdf</a>	85	Explainable AI: A survey of explainable artificial intelligence	2019	TY	

# Matching 200+ Papers with XAI Question Bank?

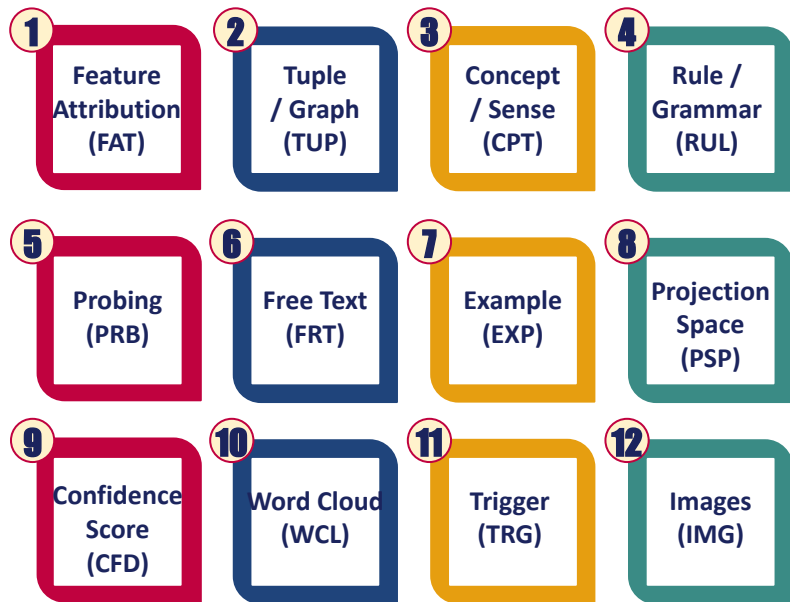
## 43 User Questions

## 218 NLP XAI Papers



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

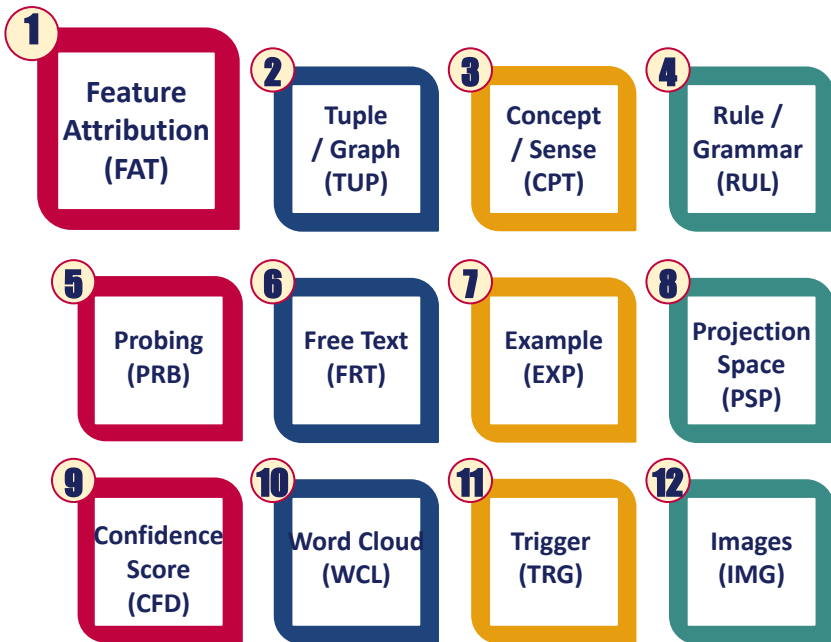
# A Collection of **XAI** Forms



## 218 NLP XAI Papers

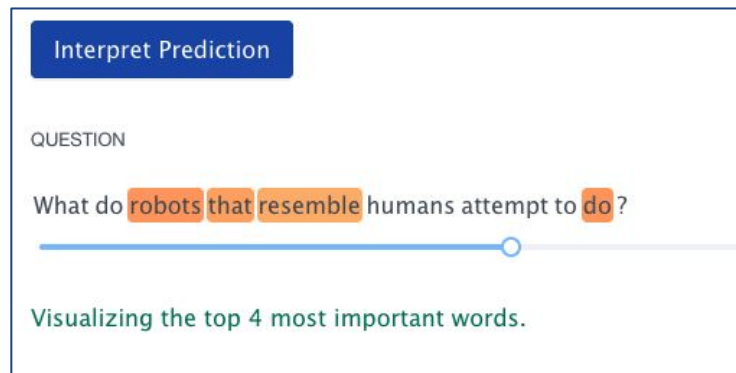
ID	Title	Year	Venue	Paper URL
1	" Why should I trust you?" Explaining the predictions of any classifier	2016	KDD	<a href="https://arxiv.org/pdf/1602.04938">https://arxiv.org/pdf/1602.04938</a>
2	Visualizing and Understanding Neural Models in NLP	2016	NAACL	<a href="https://www.aclweb.org/antholog">https://www.aclweb.org/antholog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

# 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?”





# Matching Each User Question with XAI Forms

## XAI Question Bank

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

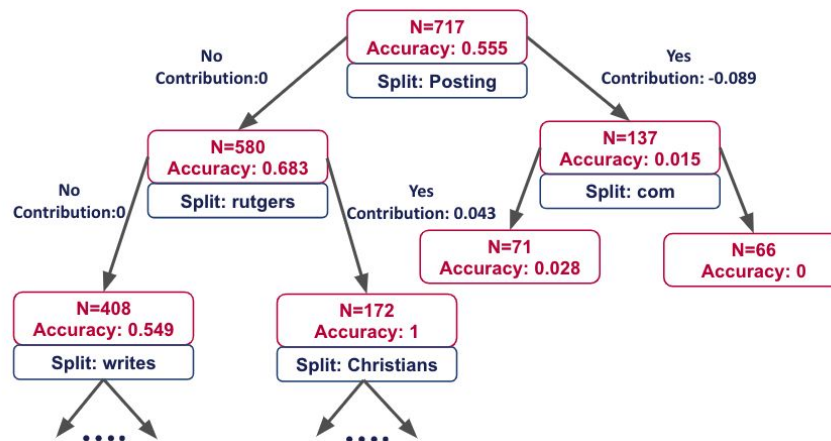
# Matching Each User Question with XAI Forms

## XAI Question Bank

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

2

Tuple  
/ Graph  
(TUP)



# Matching Each User Question with XAI Forms

## XAI Question Bank

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

4

Rule /  
Grammar  
(RUL)

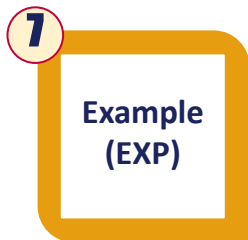
if {"not", "bad"} in input:  
then **Positive**

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

# Matching Each User Question with XAI Forms

## XAI Question Bank

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



### 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.....

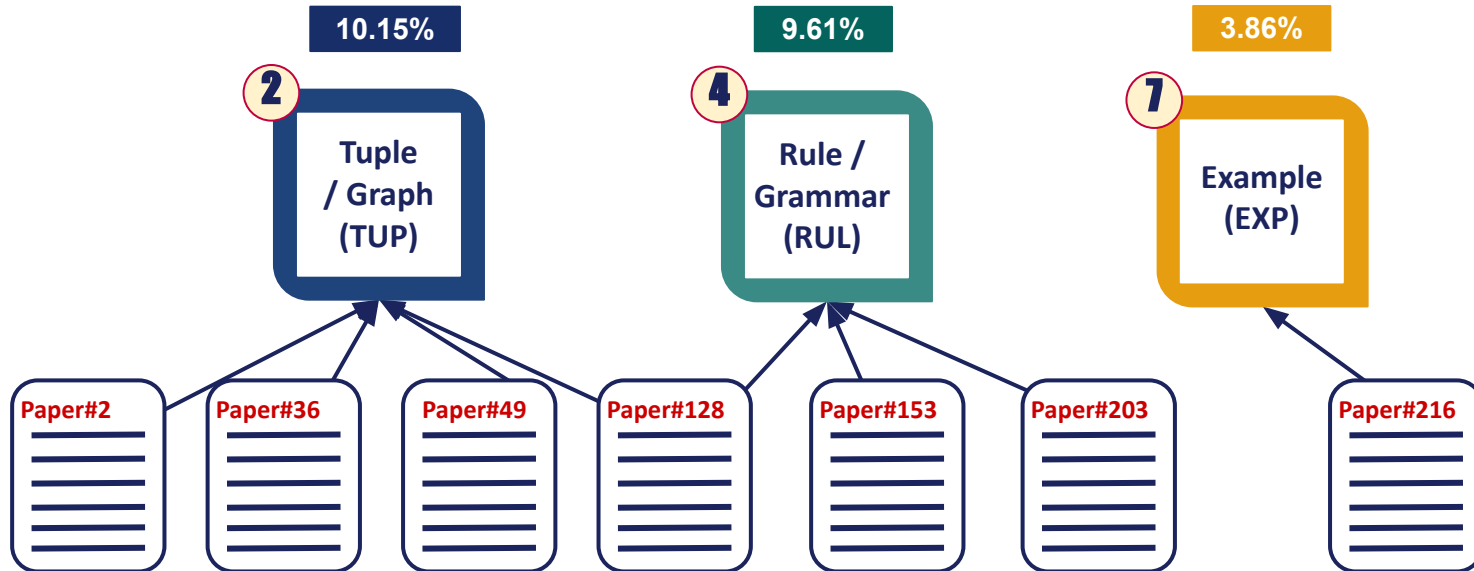
### Negative Training Examples:

- 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.....

# 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)

1

Example  
(EXP)

# Findings

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

0.0%



100.0%

# Findings

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

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

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	AAACL-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