XNLP: eXplainable NLP

Ashutosh Modi

eXplainable AI Summer School

TU Delft
August 29 – September 02, 2022

Exploration Lab
अन्वेषण प्रयोगशाला

IIT KANPUR
Indian Institute of Technology, Kanpur

Explainability Motivations

- Interpretability/Explainability: “Ability to explain or to present in understandable terms to a human” (Doshi-Velez and Kim, 2017).

Towards A Rigorous Science of Interpretable Machine Learning, Doshi-Velez and Kim. 2017, http://arxiv.org/abs/1702.08608

Explainability Motivations

- Interpretability/Explainability: “Ability to explain or to present in understandable terms to a human” (Doshi-Velez and Kim, 2017).

- Motivations:
  - Accountability

Explainability Motivations

- Interpretability/Explainability: “Ability to explain or to present in understandable terms to a human” (Doshi-Velez and Kim, 2017).

- Motivations:
  - Accountability
  - Ethics

Explainability Motivations

- Interpretability/Explainability: “Ability to explain or to present in understandable terms to a human” (Doshi-Velez and Kim, 2017).

- Motivations:
  - Accountability
  - Ethics
  - Safety and Trust

Explainability Motivations

- Interpretability/Explainability: “Ability to explain or to present in understandable terms to a human” (Doshi-Velez and Kim, 2017).

- Motivations:
  - Accountability
  - Ethics
  - Safety and Trust
  - Scientific Understanding

Measuring Explainability

Towards A Rigorous Science of Interpretable Machine Learning, Doshi-Velez and Kim. 2017, http://arxiv.org/abs/1702.08608

Post-hoc Interpretability for Neural NLP: A Survey, Madsen, 2022 (https://arxiv.org/abs/2108.04840)

Measuring Explainability

- Application Grounded

Towards A Rigorous Science of Interpretable Machine Learning, Doshi-Velez and Kim. 2017, http://arxiv.org/abs/1702.08608

Post-hoc Interpretability for Neural NLP: A Survey, Madsen, 2022 (https://arxiv.org/abs/2108.04840)

Measuring Explainability

- Application Grounded
- Human Grounded

Towards A Rigorous Science of Interpretable Machine Learning, Doshi-Velez and Kim. 2017, http://arxiv.org/abs/1702.08608

Post-hoc Interpretability for Neural NLP: A Survey, Madsen, 2022 (https://arxiv.org/abs/2108.04840)

Measuring Explainability

- Application Grounded
- Human Grounded
- Functionally Grounded

Towards A Rigorous Science of Interpretable Machine Learning, Doshi-Velez and Kim. 2017, http://arxiv.org/abs/1702.08608

Post-hoc Interpretability for Neural NLP: A Survey, Madsen, 2022 (https://arxiv.org/abs/2108.04840)

Measuring Explainability

- Application Grounded
- Human Grounded
- Functionally Grounded

Practical Approach:

Compare with intrinsically interpretable model

Propose desirable

Benchmark against random explanations

Towards A Rigorous Science of Interpretable Machine Learning, Doshi-Velez and Kim. 2017, http://arxiv.org/abs/1702.08608

Post-hoc Interpretability for Neural NLP: A Survey, Madsen, 2022 (https://arxiv.org/abs/2108.04840)

Typical NLP Problems/Tasks

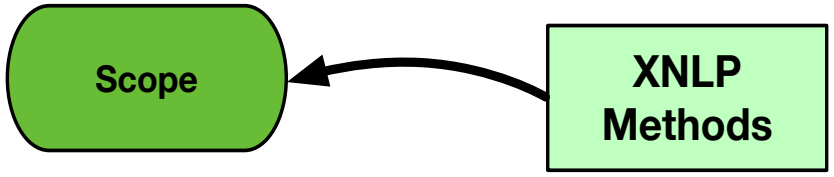
- Sequence to class prediction tasks
- Sequence to sequence prediction tasks

XNLP Methods Perspectives

XNLP Methods

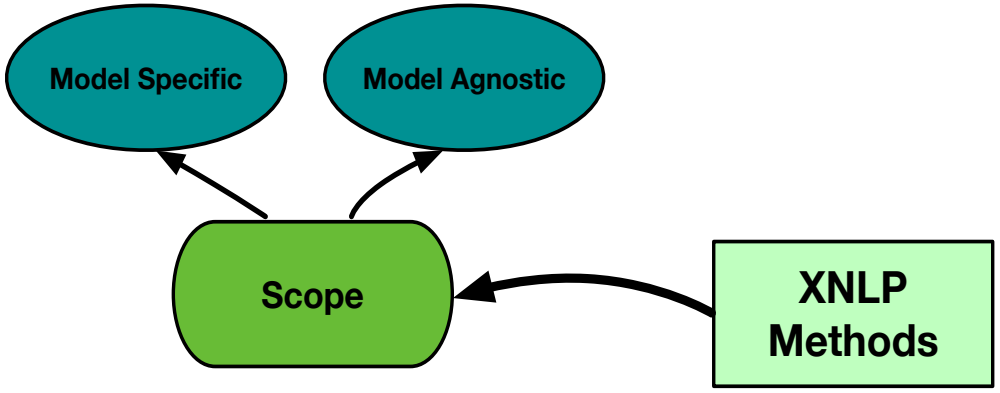
Post-hoc Interpretability for Neural NLP: A Survey, Madsen, 2022 (https://arxiv.org/abs/2108.04840)
Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book/

XNLP Methods Perspectives



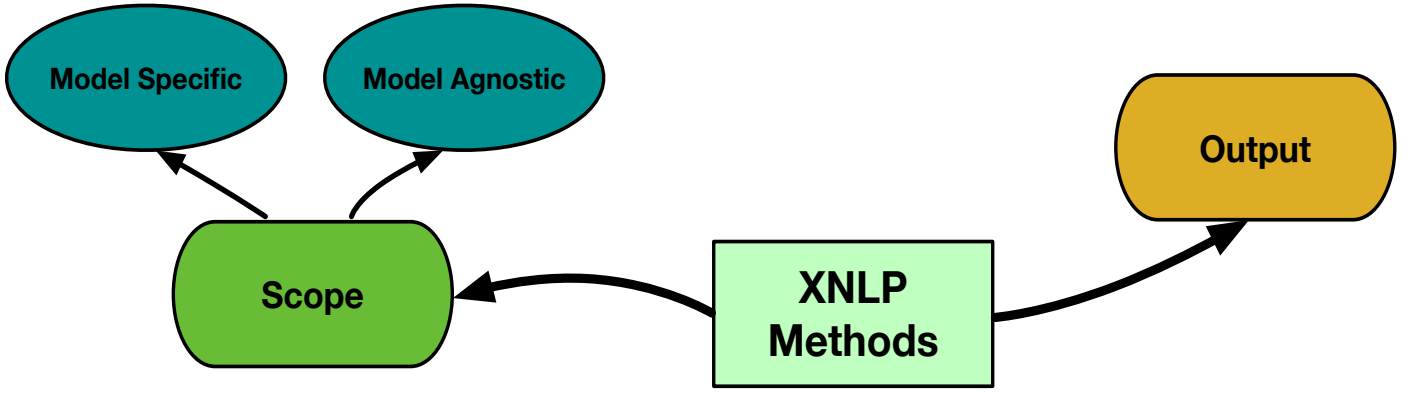
Post-hoc Interpretability for Neural NLP: A Survey, Madsen, 2022 (https://arxiv.org/abs/2108.04840)
Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book/

XNLP Methods Perspectives



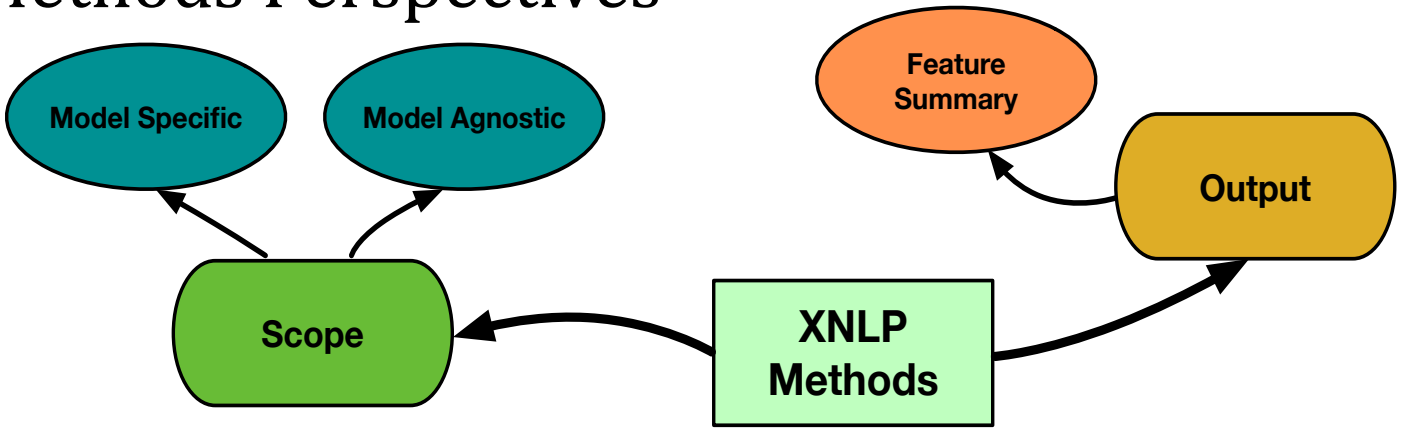
Post-hoc Interpretability for Neural NLP: A Survey, Madsen, 2022 (https://arxiv.org/abs/2108.04840)
Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book/

XNLP Methods Perspectives



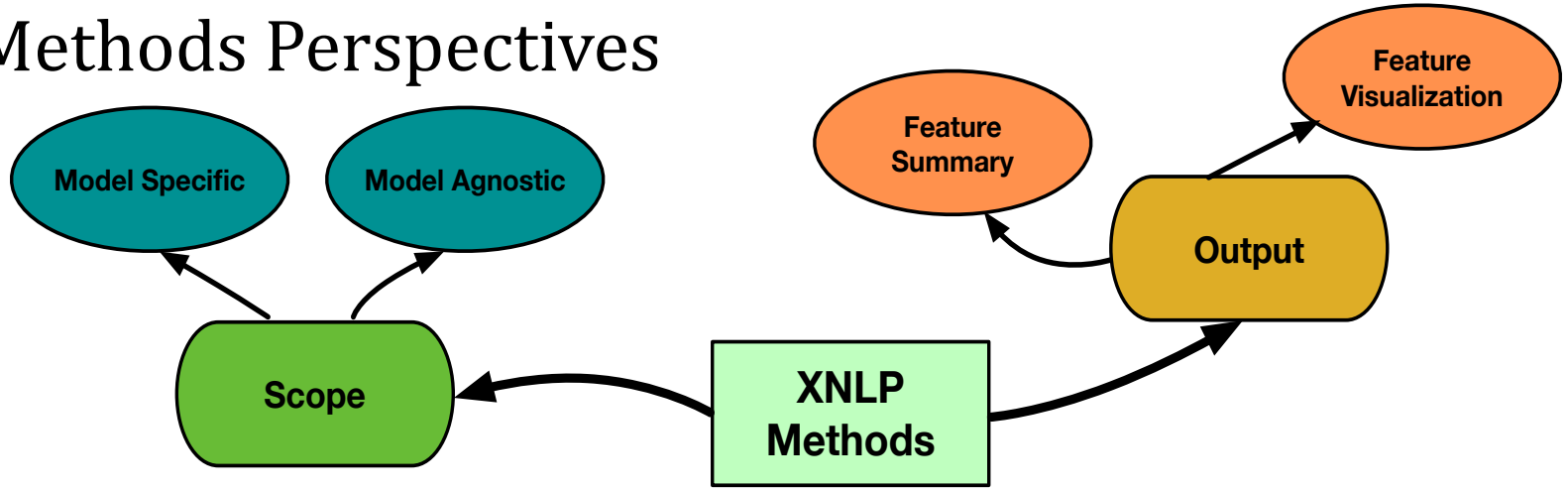
Post-hoc Interpretability for Neural NLP: A Survey, Madsen, 2022 (https://arxiv.org/abs/2108.04840)
Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book/

XNLP Methods Perspectives



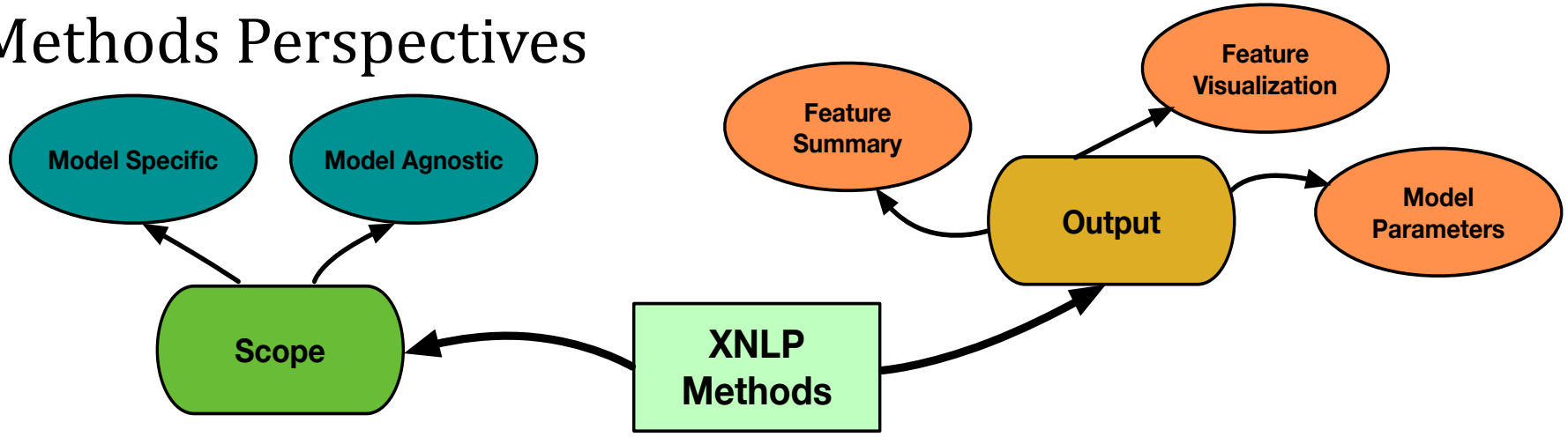
Post-hoc Interpretability for Neural NLP: A Survey, Madsen, 2022 (https://arxiv.org/abs/2108.04840)
Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book/

XNLP Methods Perspectives



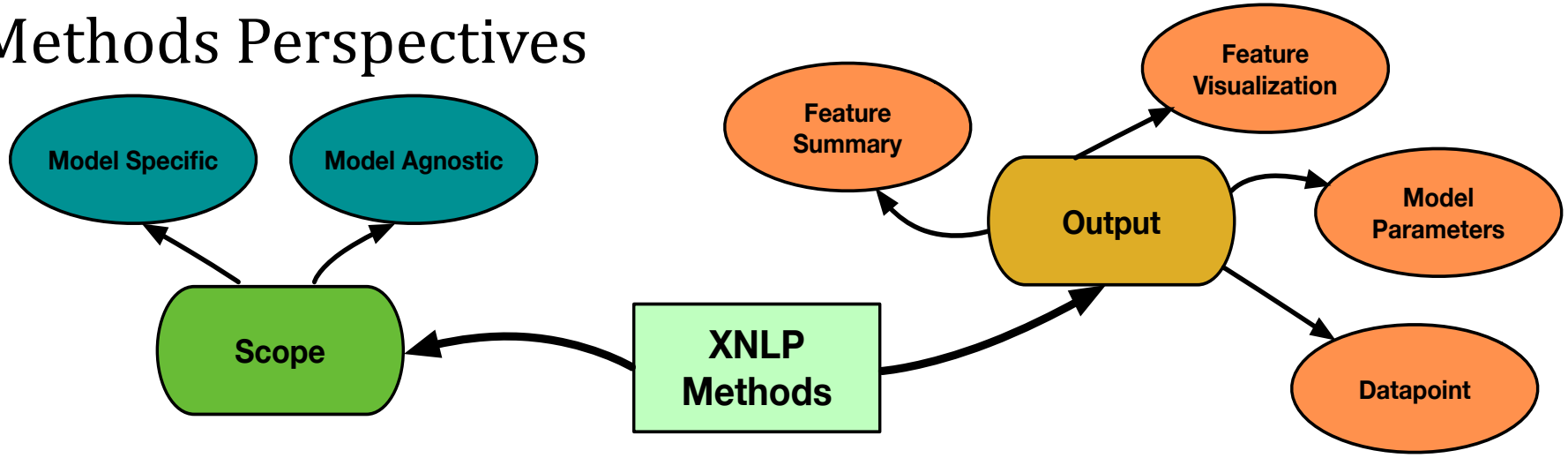
Post-hoc Interpretability for Neural NLP: A Survey, Madsen, 2022 (https://arxiv.org/abs/2108.04840)
Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book/

XNLP Methods Perspectives



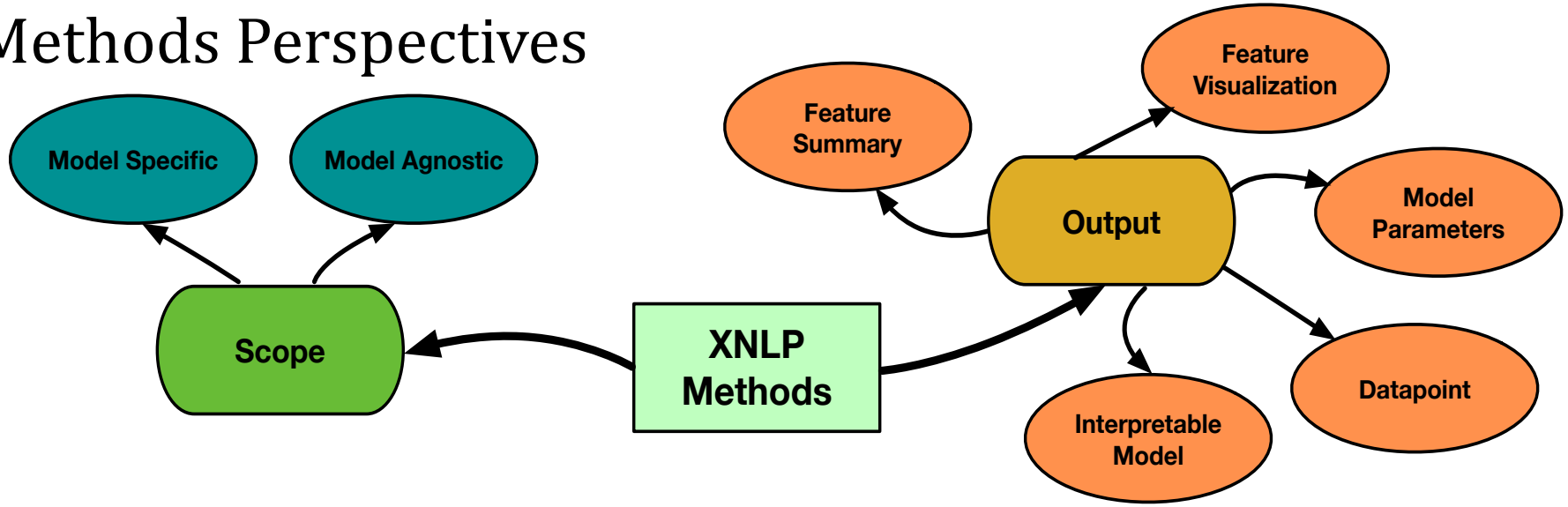
Post-hoc Interpretability for Neural NLP: A Survey, Madsen, 2022 (https://arxiv.org/abs/2108.04840)
Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book/

XNLP Methods Perspectives



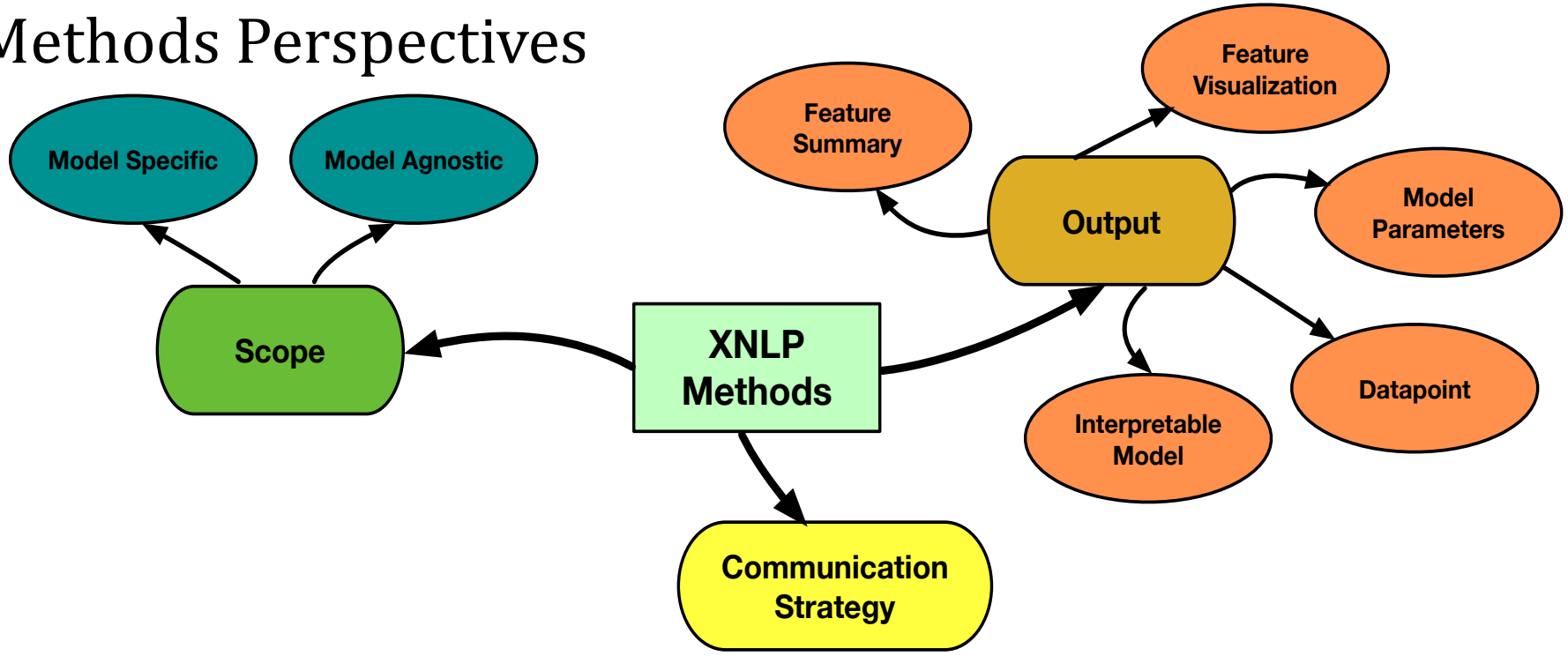
Post-hoc Interpretability for Neural NLP: A Survey, Madsen, 2022 (https://arxiv.org/abs/2108.04840)
Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book/

XNLP Methods Perspectives



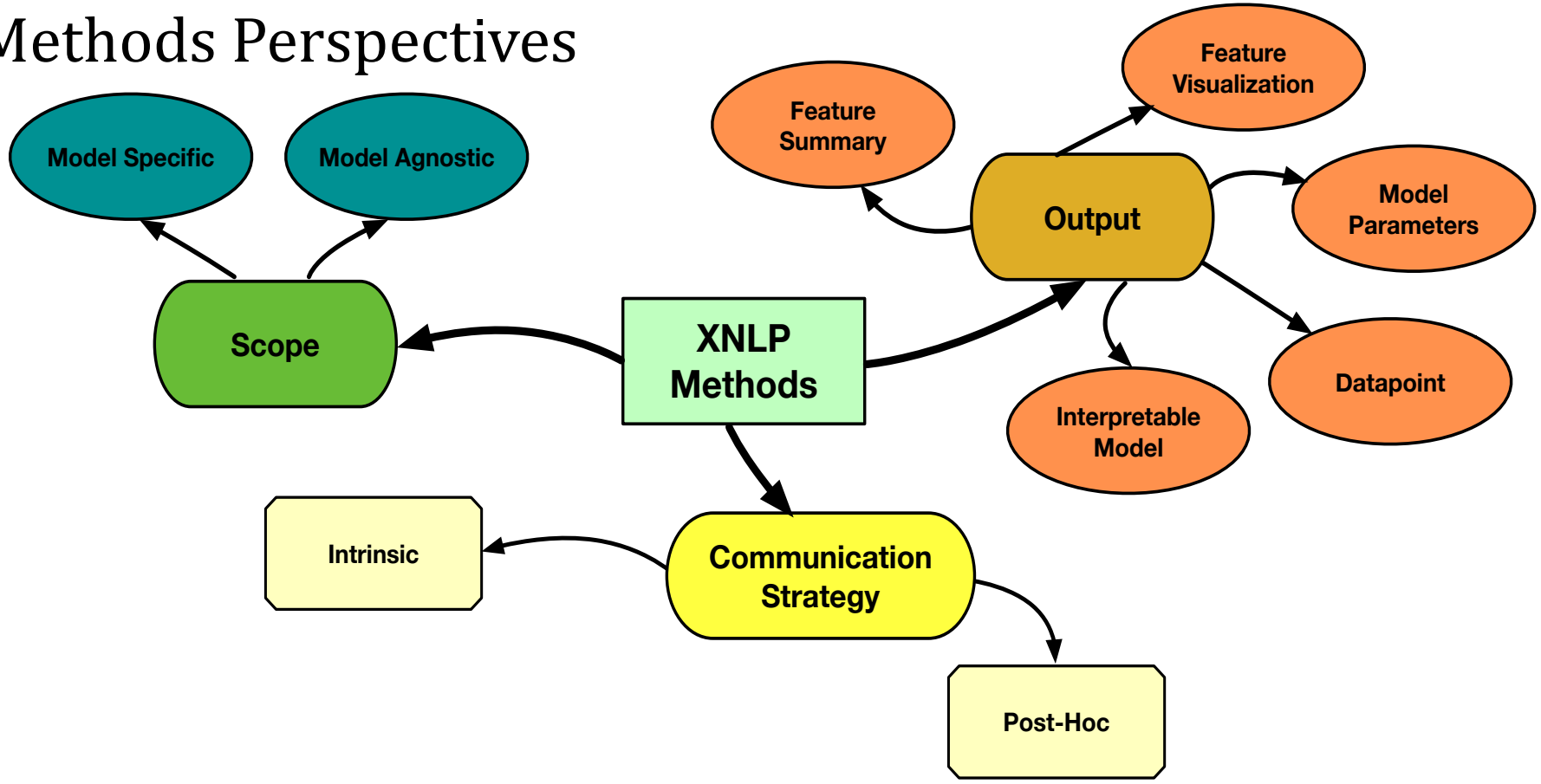
Post-hoc Interpretability for Neural NLP: A Survey, Madsen, 2022 (https://arxiv.org/abs/2108.04840)
Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book/

XNLP Methods Perspectives



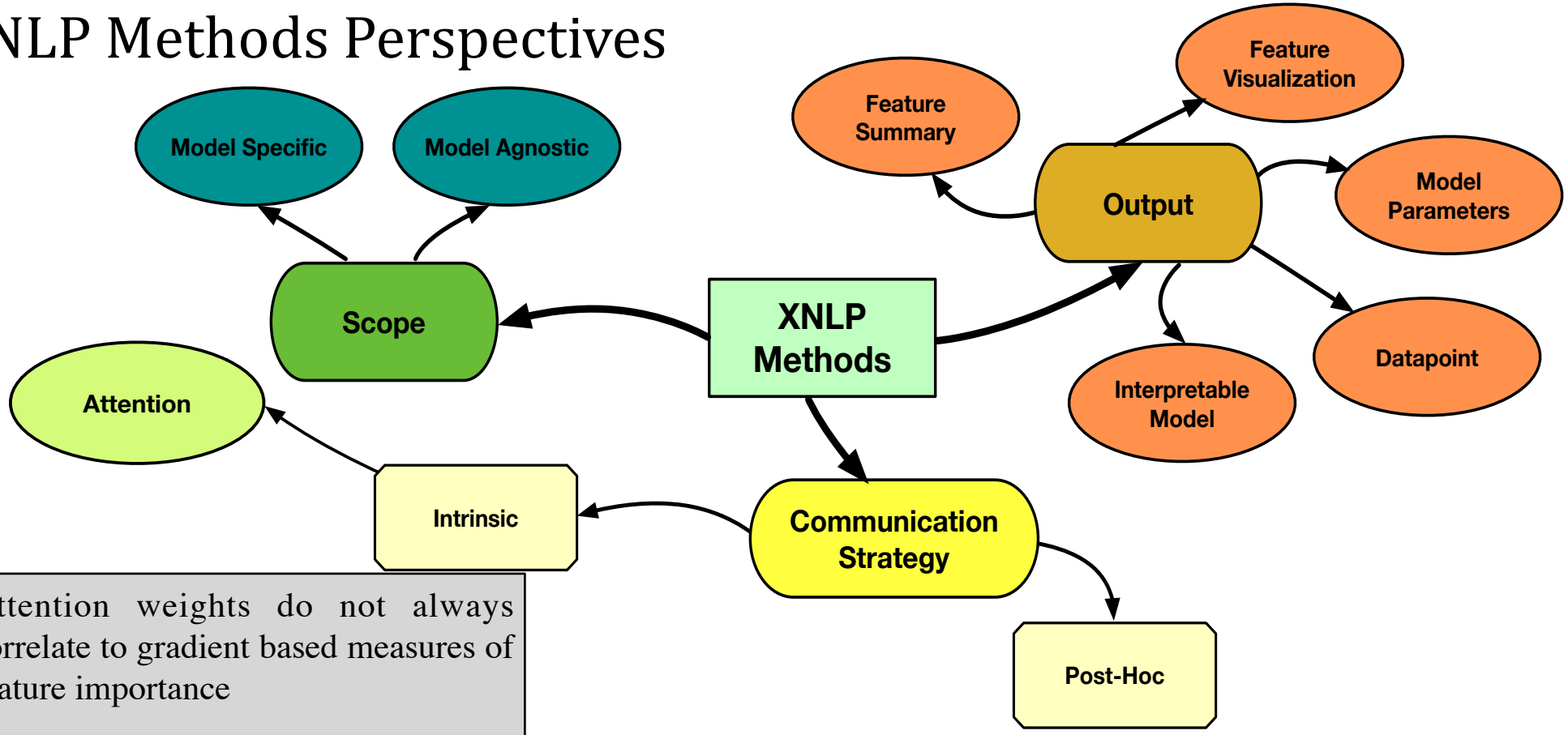
Post-hoc Interpretability for Neural NLP: A Survey, Madsen, 2022 (https://arxiv.org/abs/2108.04840)
Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book/

XNLP Methods Perspectives



Post-hoc Interpretability for Neural NLP: A Survey, Madsen, 2022 (https://arxiv.org/abs/2108.04840)
Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book/

XNLP Methods Perspectives

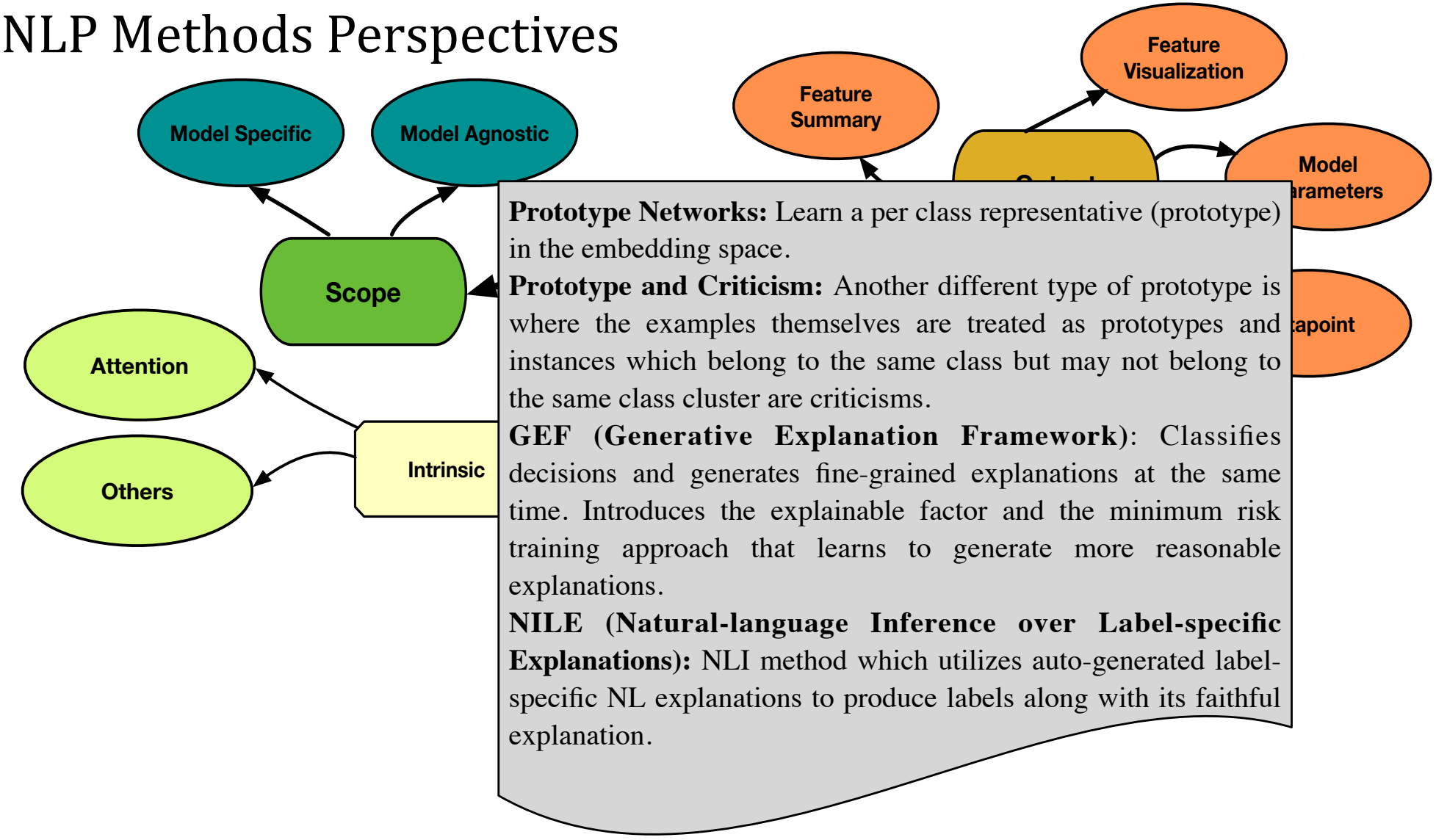


Attention weights do not always correlate to gradient based measures of feature importance

Attention distribution is not unique, for a completely different (possibly meaningless) weight distribution you could get the same prediction

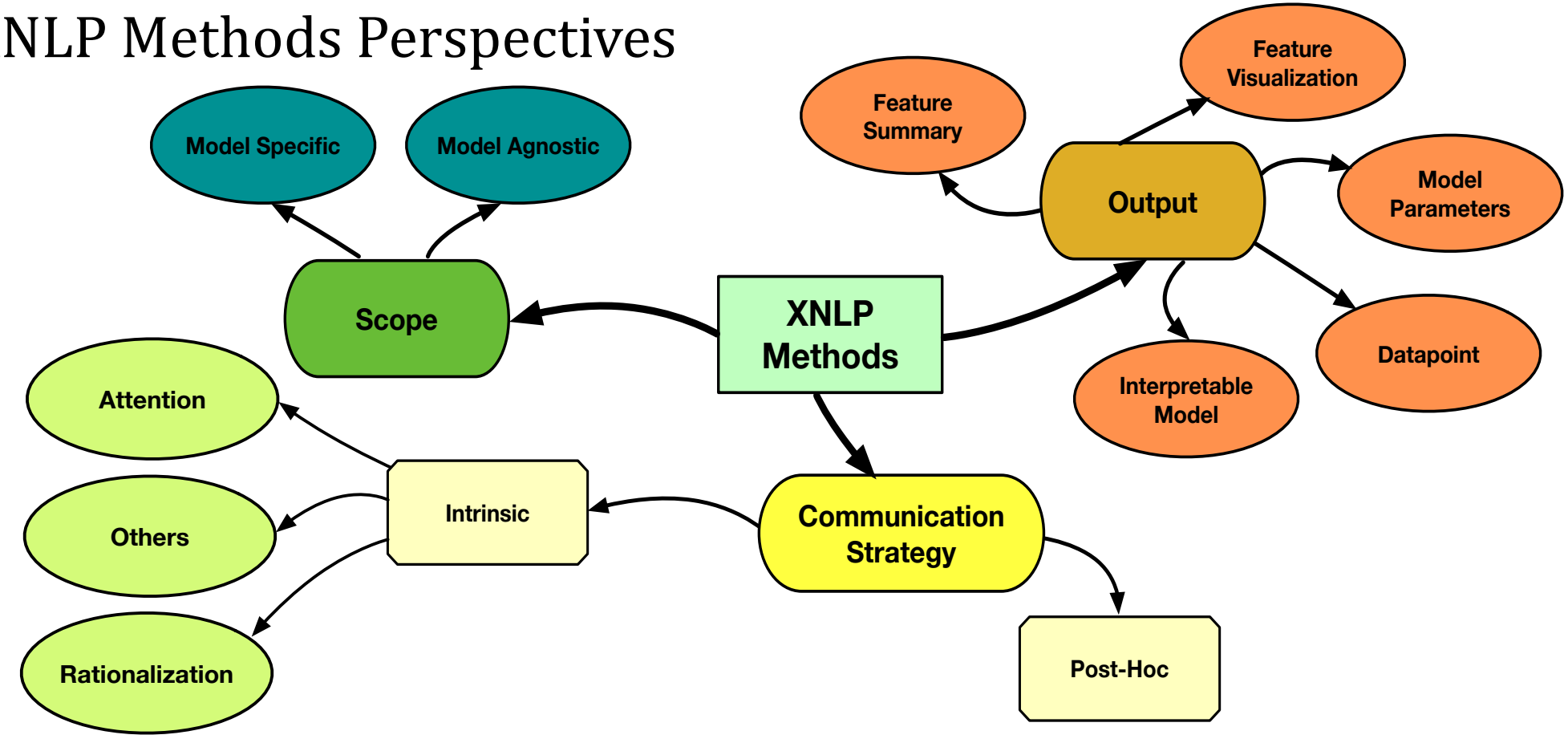
Attention is not Explanation, Jain and Wallace, NAACL, 2019, https://arxiv.org/abs/1902.10186

XNLP Methods Perspectives

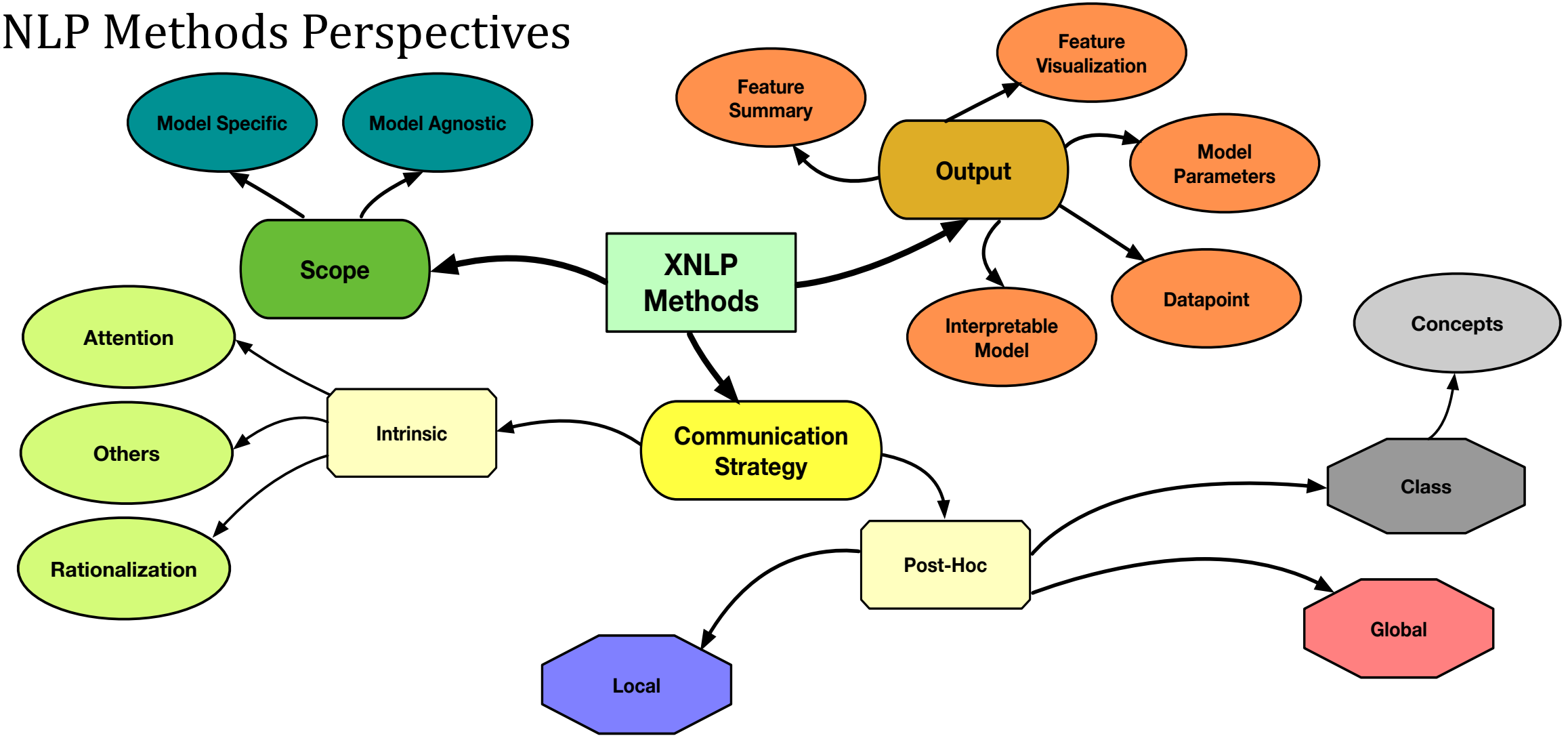


Prototypical Networks for Few-shot Learning, Snell, et al., NeurIPS, 2017, https://arxiv.org/abs/1703.05175
Examples are not enough, learn to criticize! Criticism for Interpretability, Kim, et al., NeurIPS 2016: https://tinyurl.com/2p9b8t7f
Towards Explainable NLP: A Generative Explanation Framework for Text Classification, Liu et al., ACL 2019: https://doi.org/10.18653/v1/P19-1560
NILE : Natural Language Inference with Faithful Natural Language Explanations. Kumar et al., ACL 2020: https://aclanthology.org/2020.acl-main.771/

XNLP Methods Perspectives

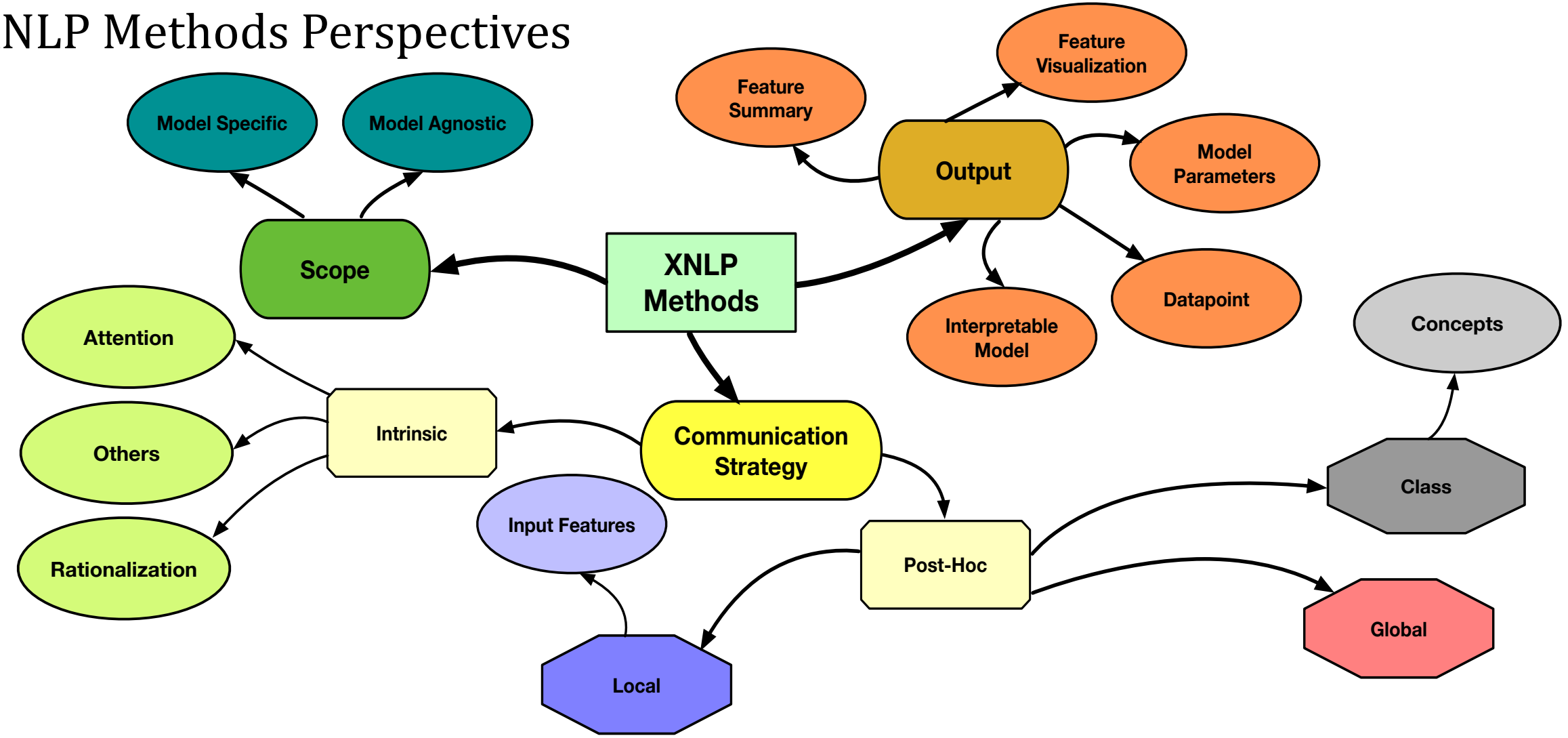


XNLP Methods Perspectives

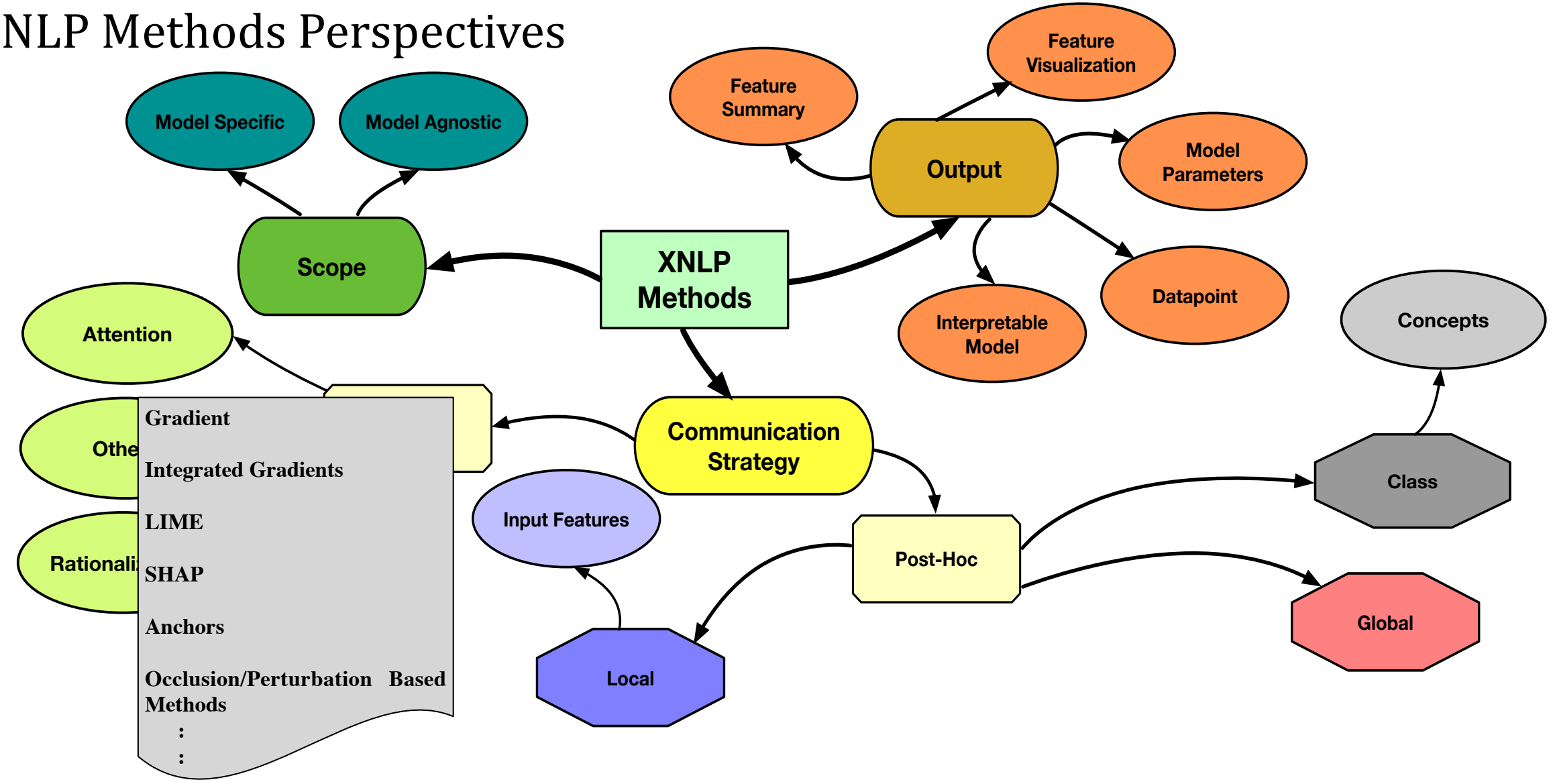


Investigating Gender Bias in Language Models Using Causal Mediation Analysis, Vig et al., 2020: https://proceedings.neurips.cc/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf

XNLP Methods Perspectives



XNLP Methods Perspectives



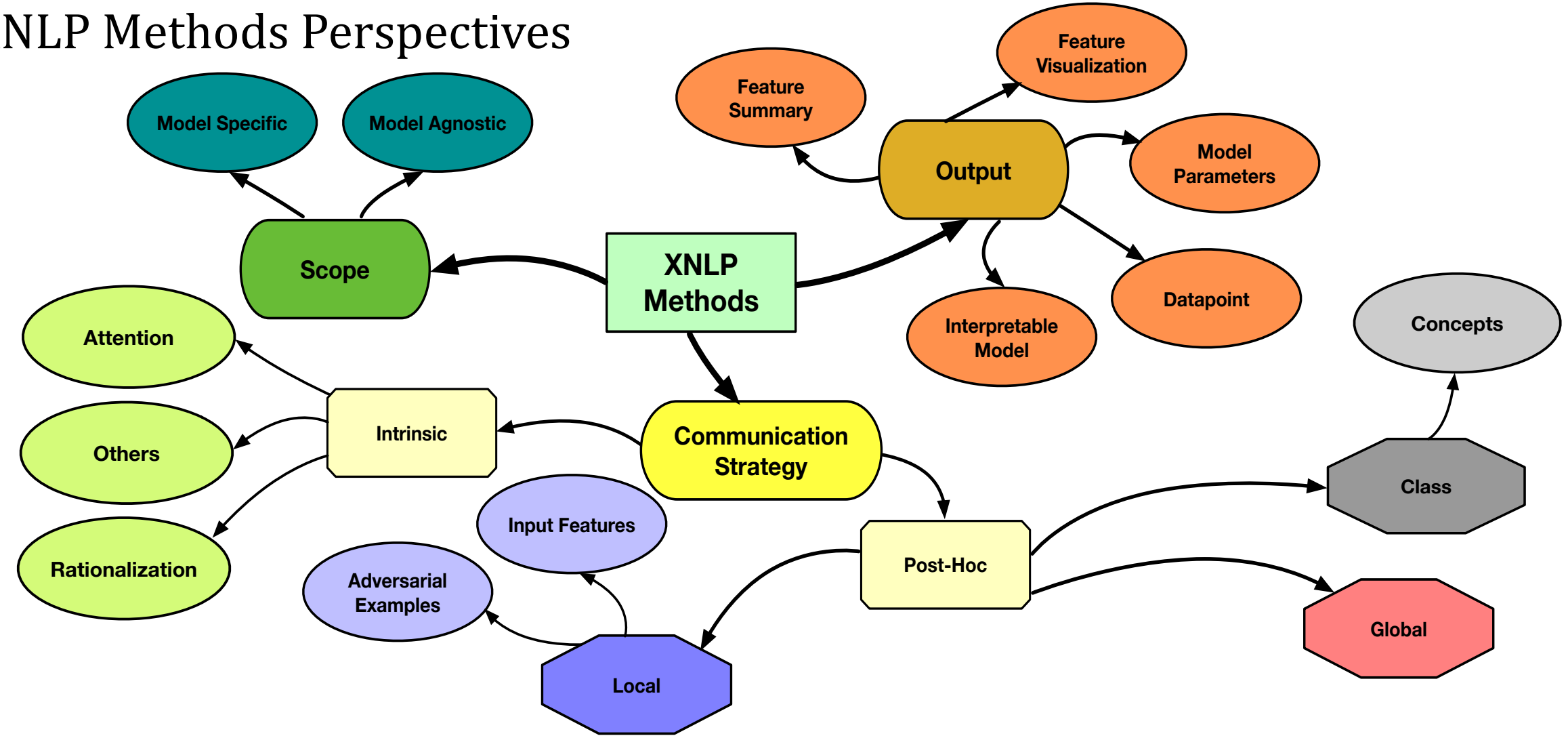
Axiomatic attribution for deep networks, Sundararajan et al., ICML 2017, http://arxiv.org/abs/1703.01365

"Why Should I Trust You?": Explaining the Predictions of Any Classifier, Riberio et al., 2016, https://dl.acm.org/doi/10.1145/2939672.2939778

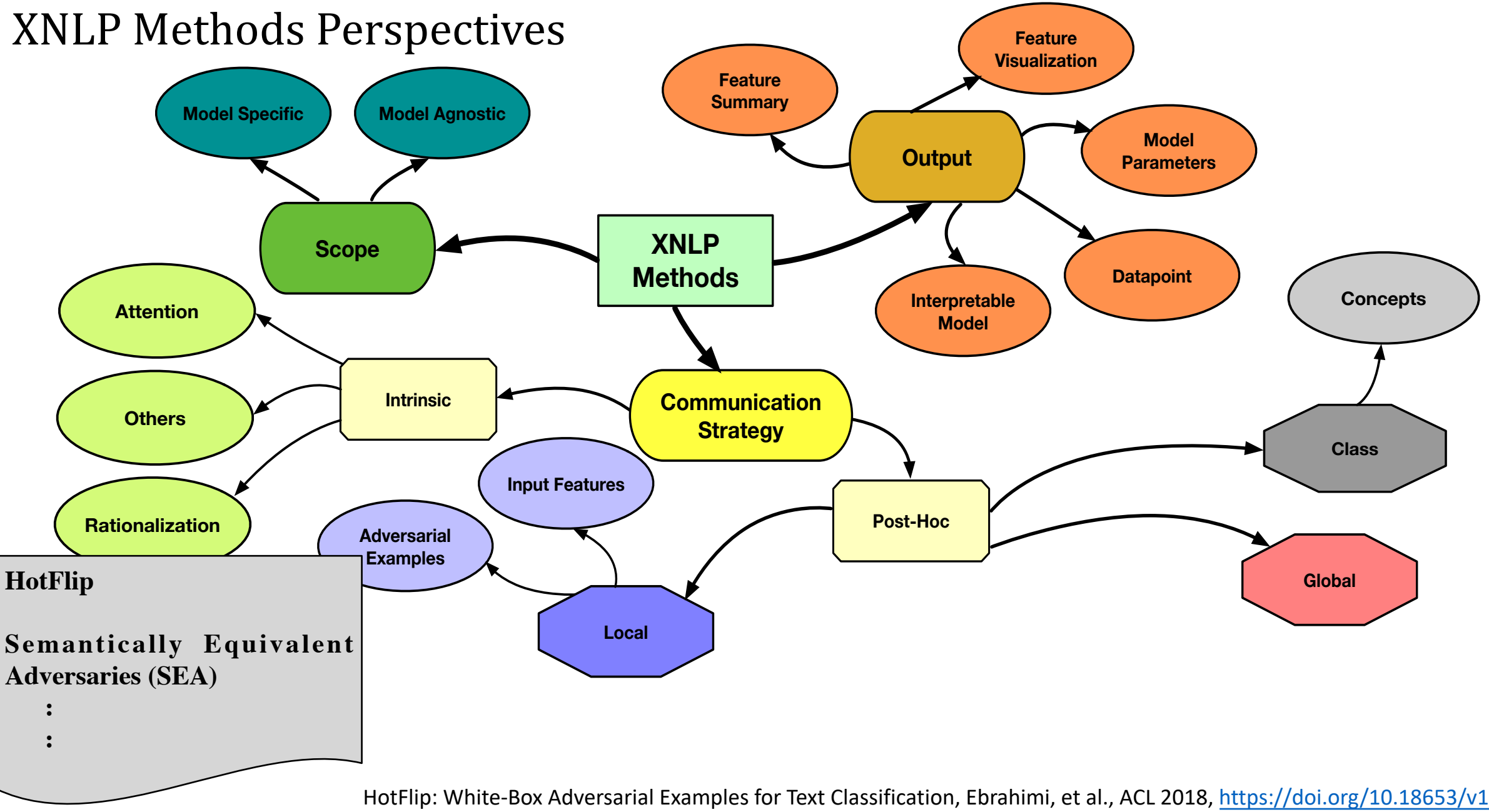
A Unified Approach to Interpreting Model Predictions, Lundberg and Lee, 2017, http://arxiv.org/abs/1705.07874

ANCHORS: High-Precision Model-Agnostic Explanations, Riberio et al., 2018, https://homes.cs.washington.edu/~marcotcr/aaai18.pdf

XNLP Methods Perspectives



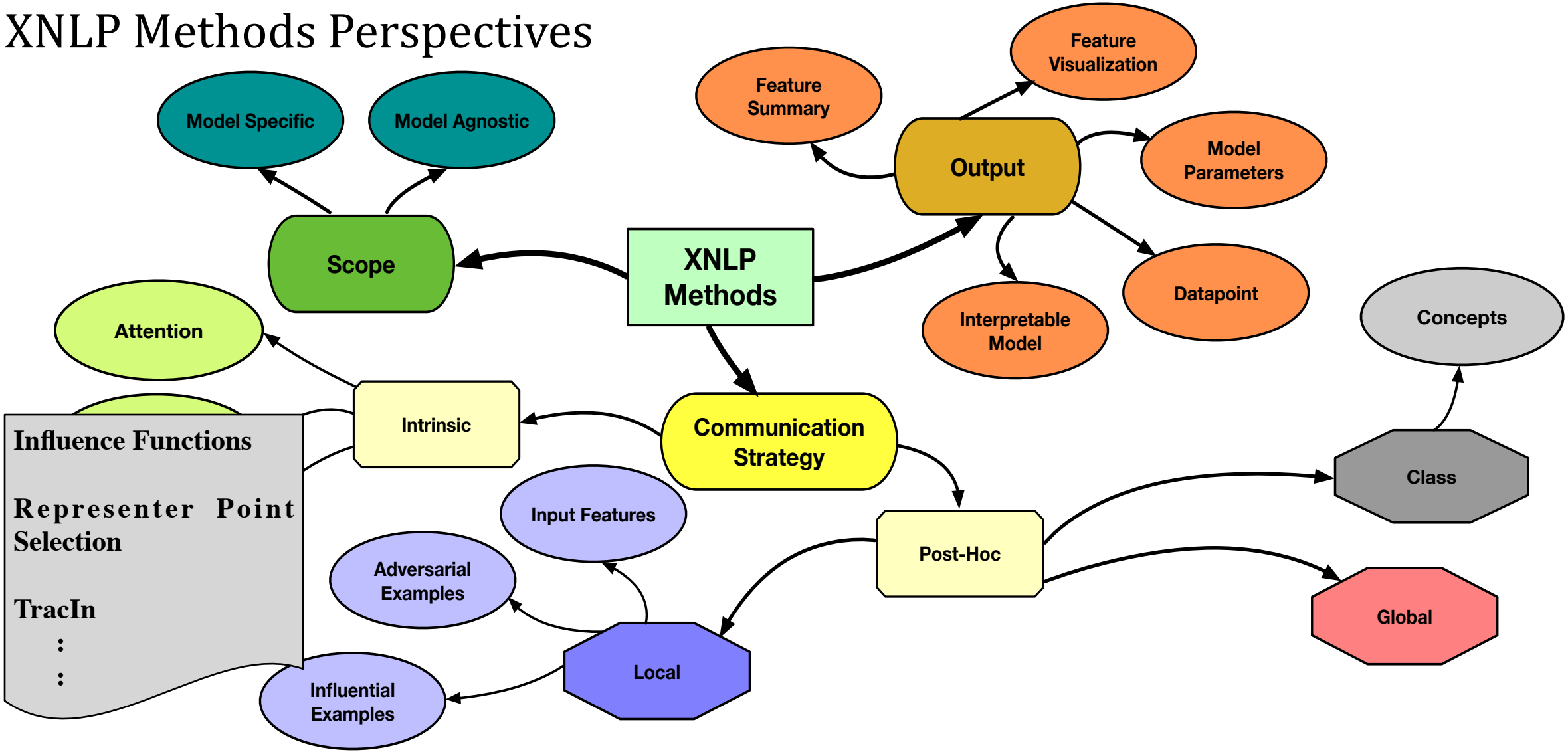
XNLP Methods Perspectives



HotFlip: White-Box Adversarial Examples for Text Classification, Ebrahimi, et al., ACL 2018, https://doi.org/10.18653/v1/P18-2006

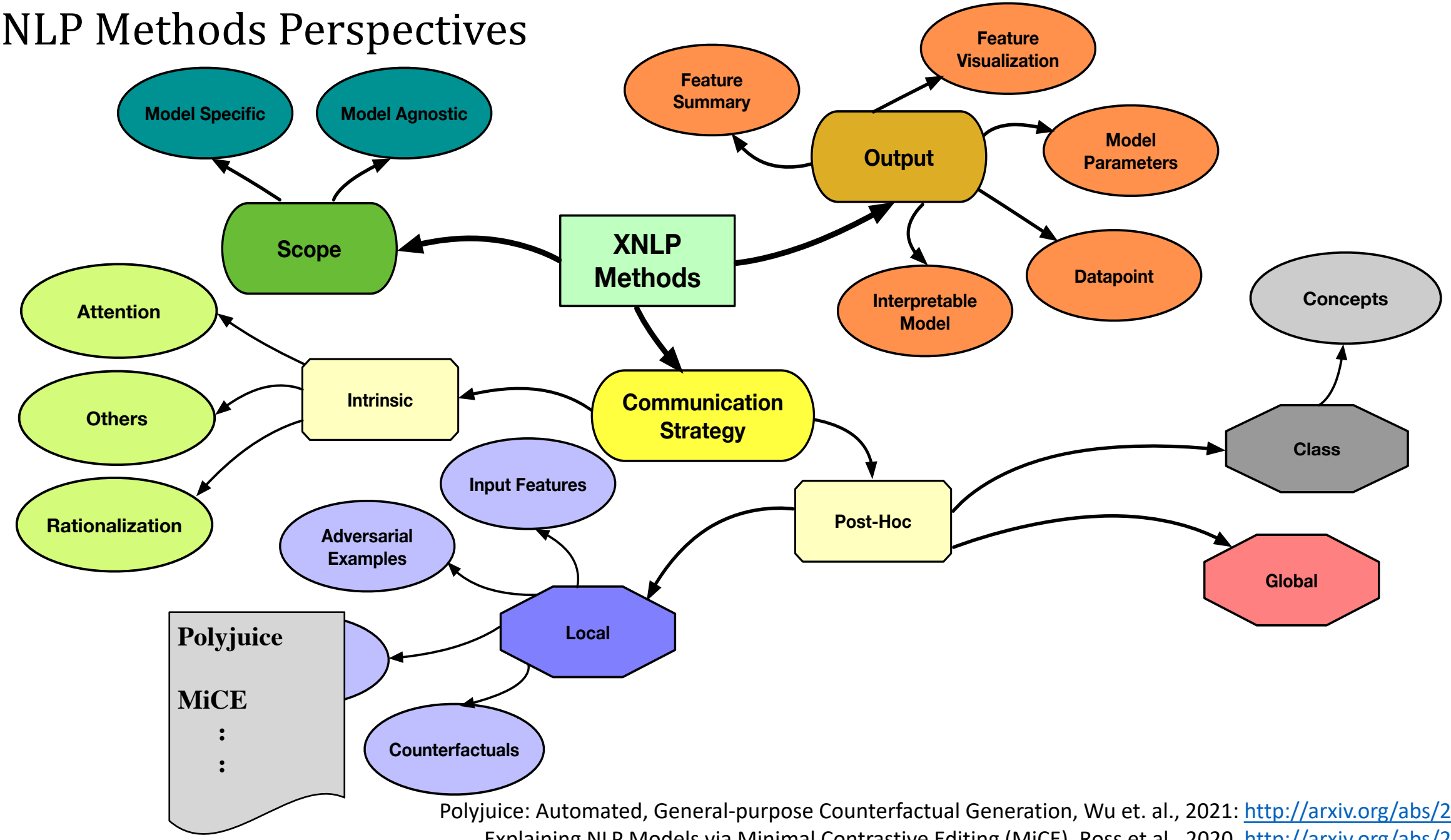
Semantically Equivalent Adversarial Rules for Debugging NLP models, Ribeiro, et al., ACL. 2018, https://doi.org/10.18653/v1/P18-1079

XNLP Methods Perspectives



Understanding Black-box Predictions via Influence Functions, Koh and Liang, ICML 2017, http://arxiv.org/abs/1703.04730
Representer Point Selection for Explaining Deep Neural Networks, Yeh, et al., 2018, http://arxiv.org/abs/1811.09720
Estimating Training Data Influence by Tracing Gradient Descent, Pruthi et al., 2020, http://arxiv.org/abs/2002.08484

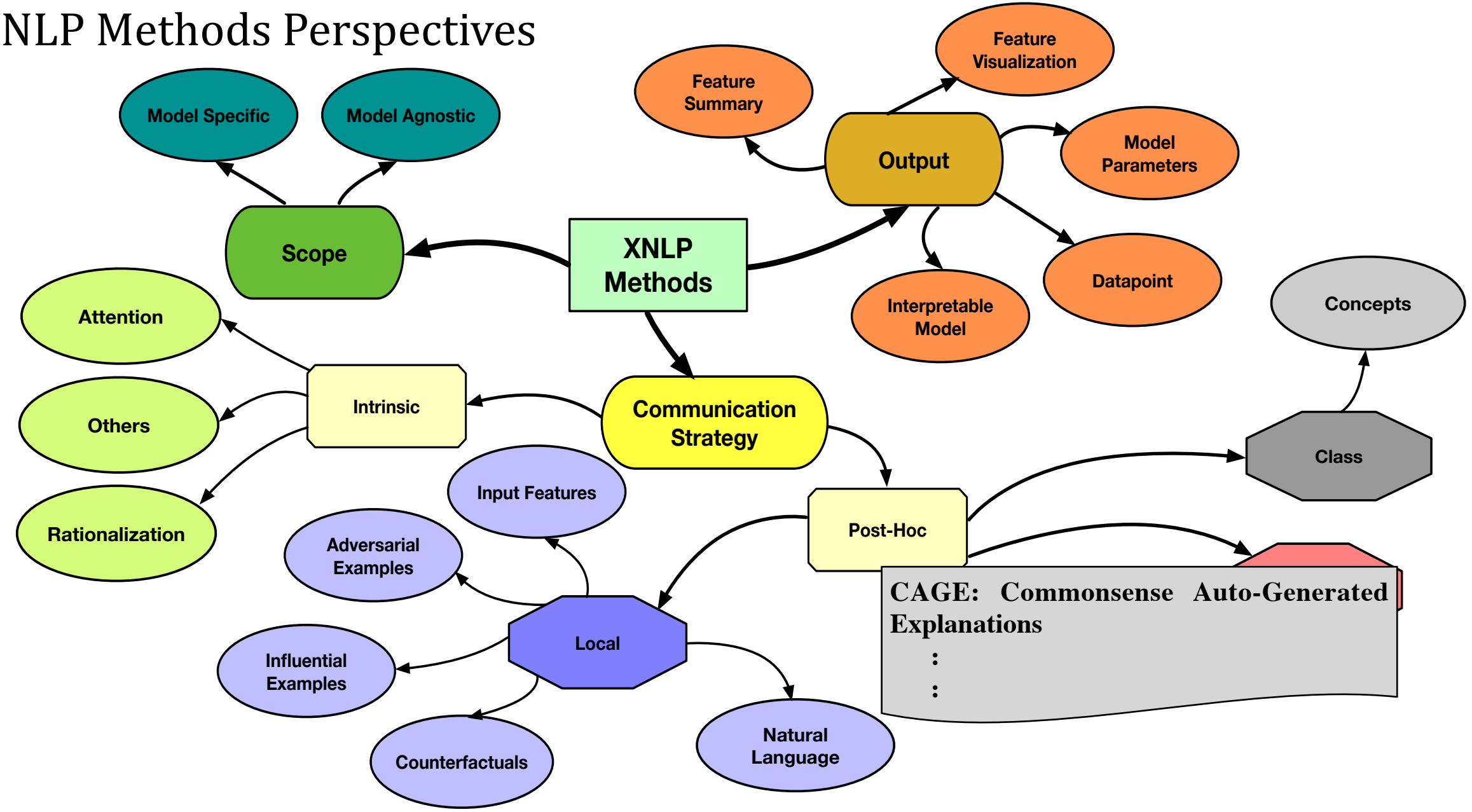
XNLP Methods Perspectives



Polyjuice: Automated, General-purpose Counterfactual Generation, Wu et. al., 2021: http://arxiv.org/abs/2101.00288

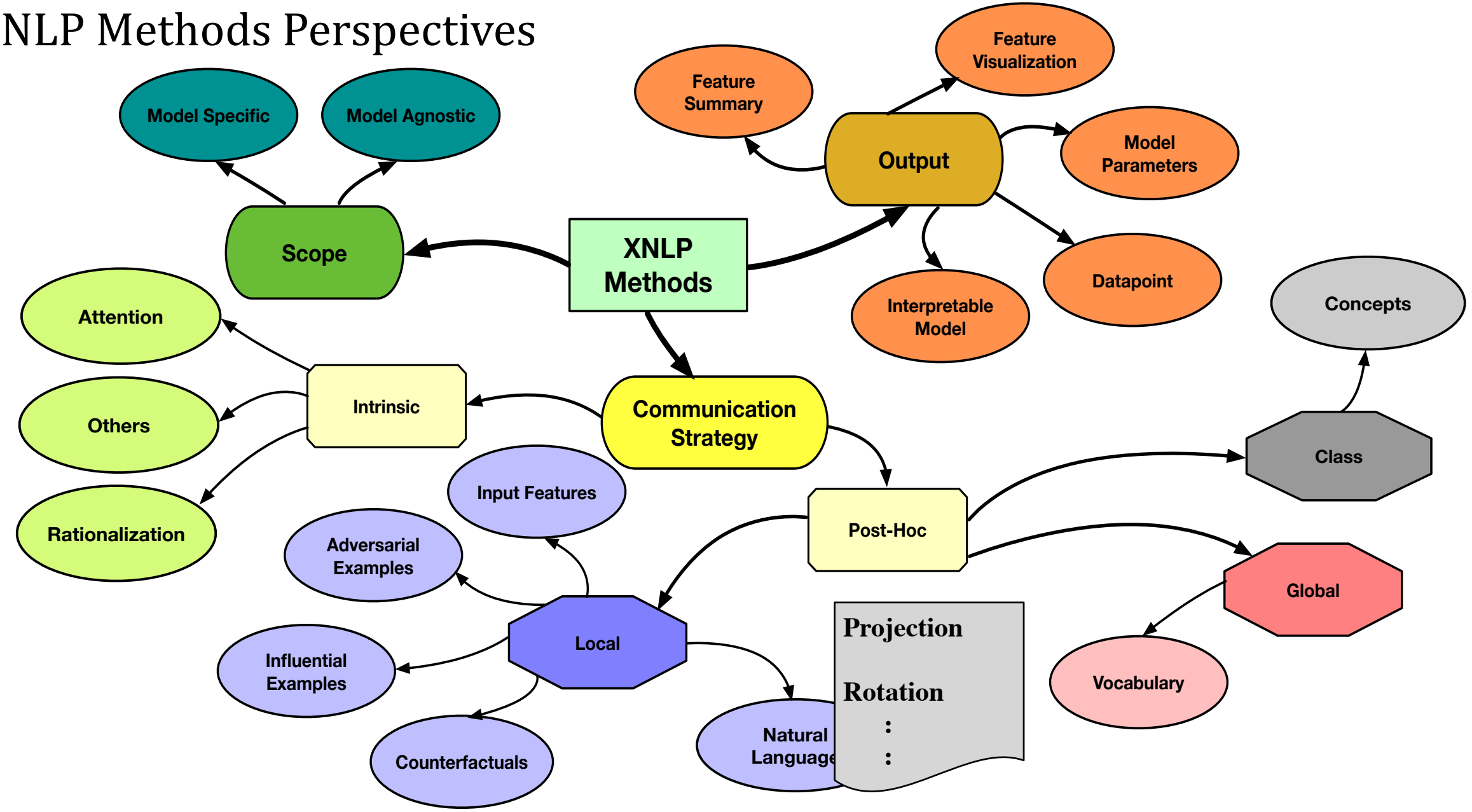
Explaining NLP Models via Minimal Contrastive Editing (MiCE), Ross et al., 2020, http://arxiv.org/abs/2012.13985

XNLP Methods Perspectives



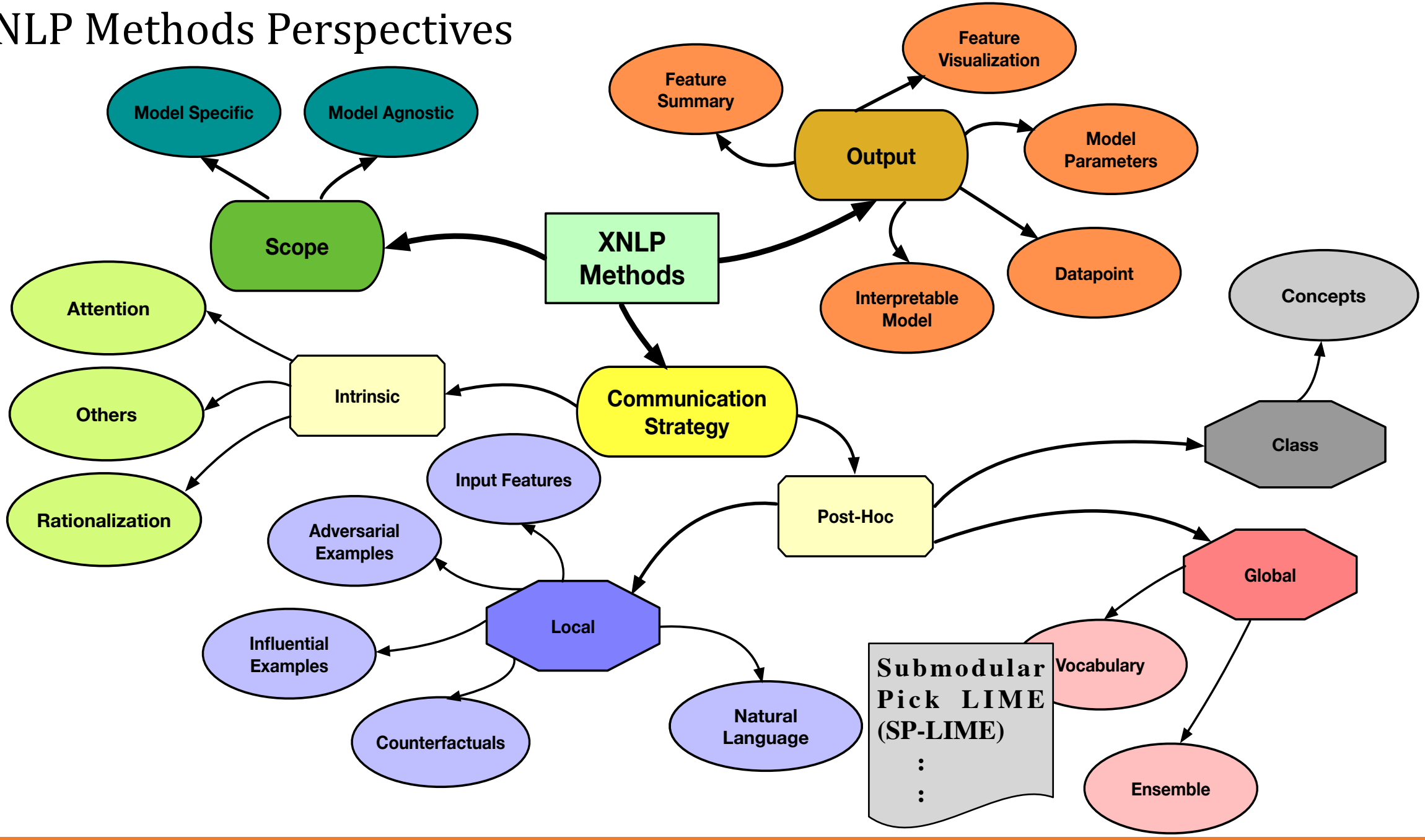
Explain Yourself! Leveraging Language Models for Commonsense Reasoning, Rajani, et al., 2019, https://doi.org/10.18653/v1/P19-1487

XNLP Methods Perspectives

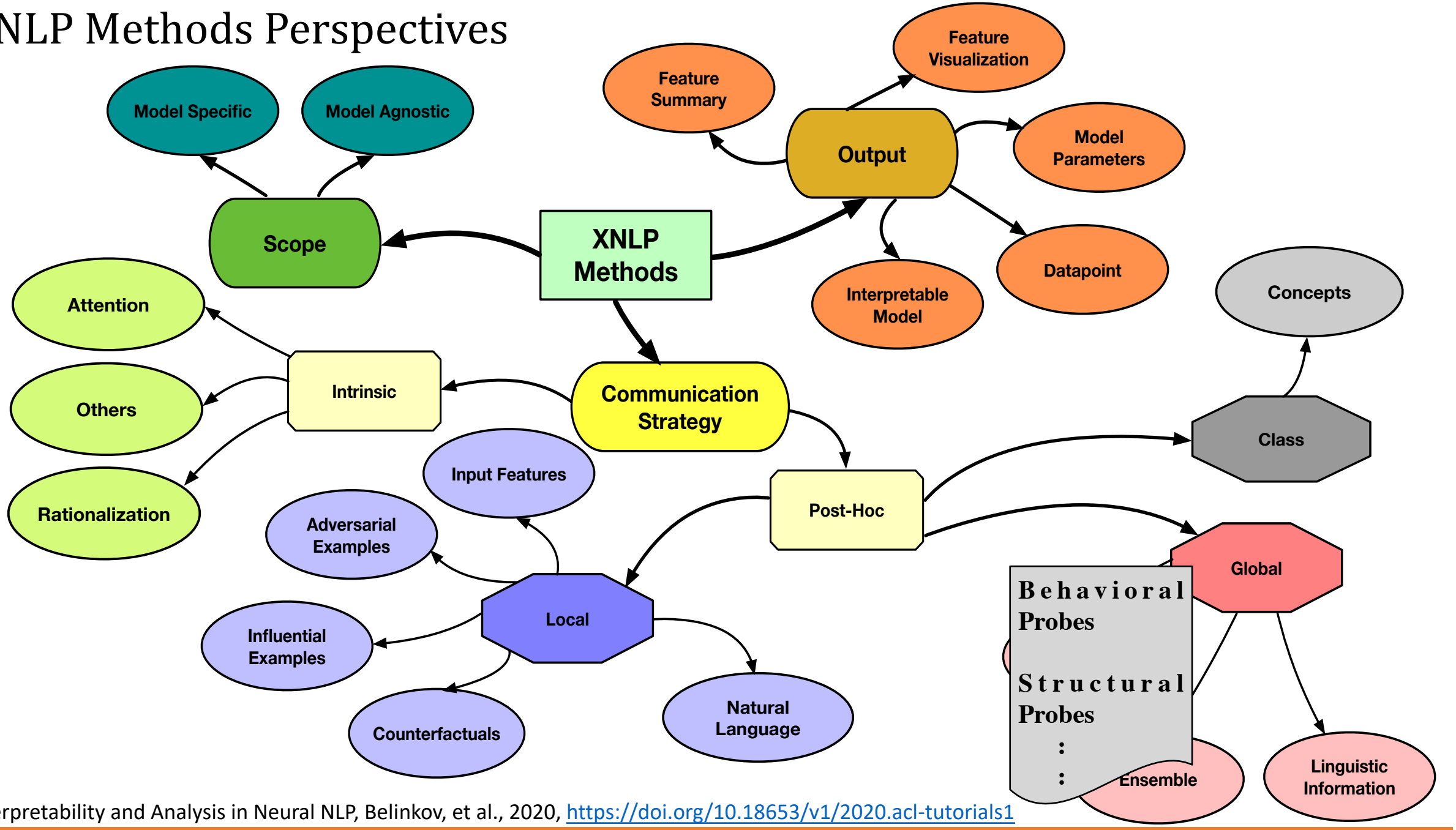


Park, et al., 2017. Rotated Word Vector Representations and their Interpretability: https://doi.org/10.18653/v1/D17-1041

XNLP Methods Perspectives

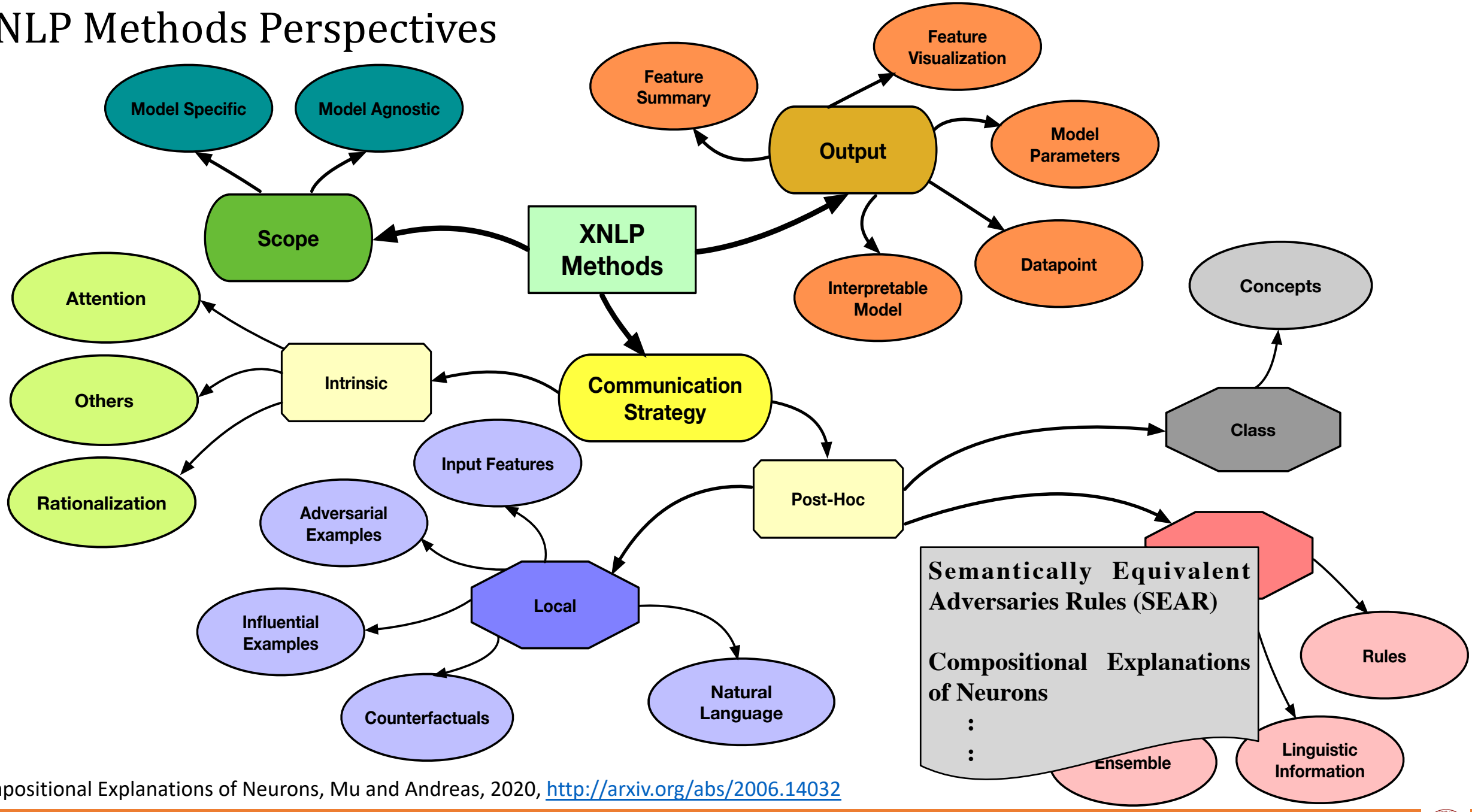


XNLP Methods Perspectives



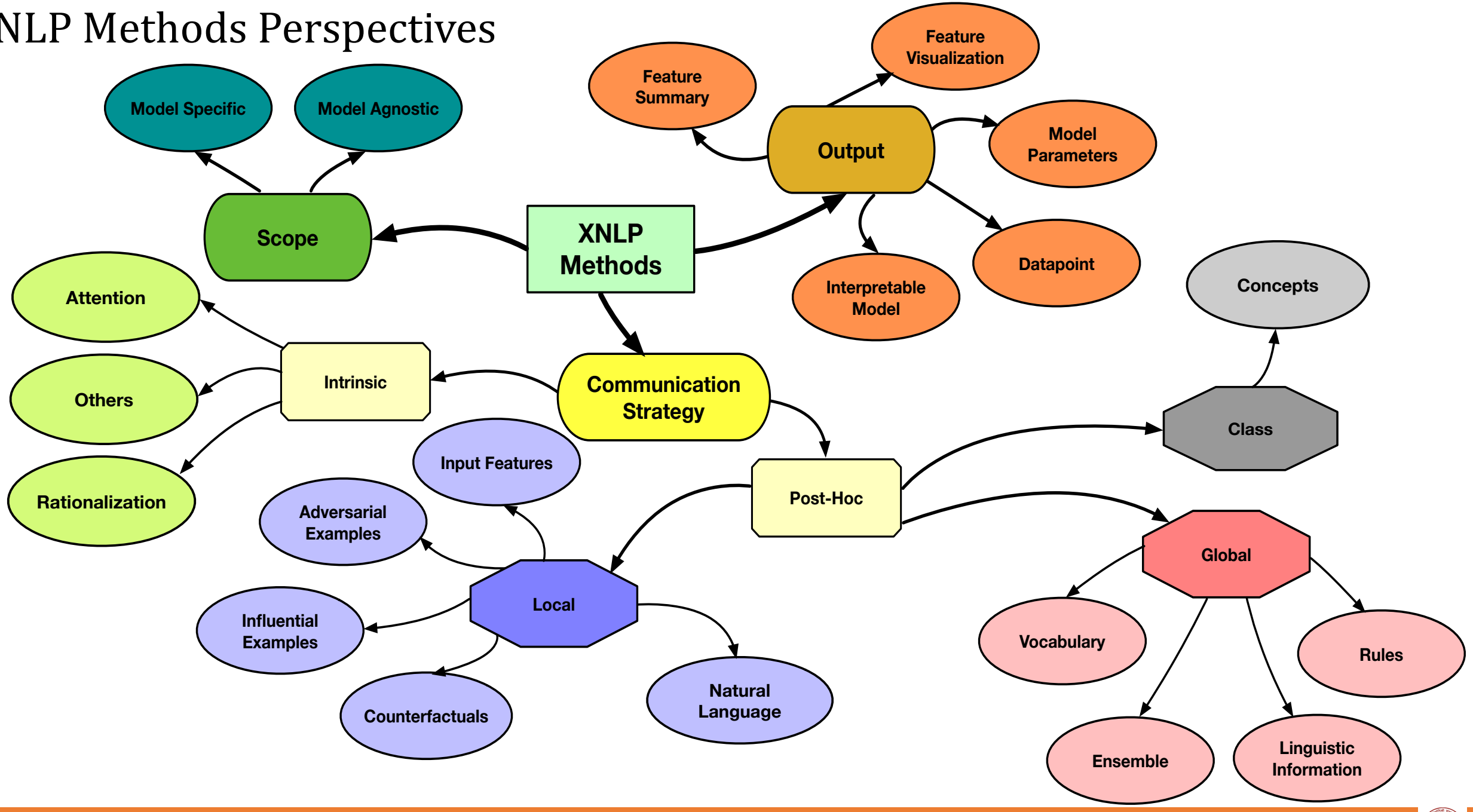
Interpretability and Analysis in Neural NLP, Belinkov, et al., 2020, https://doi.org/10.18653/v1/2020.acl-tutorials1

XNLP Methods Perspectives

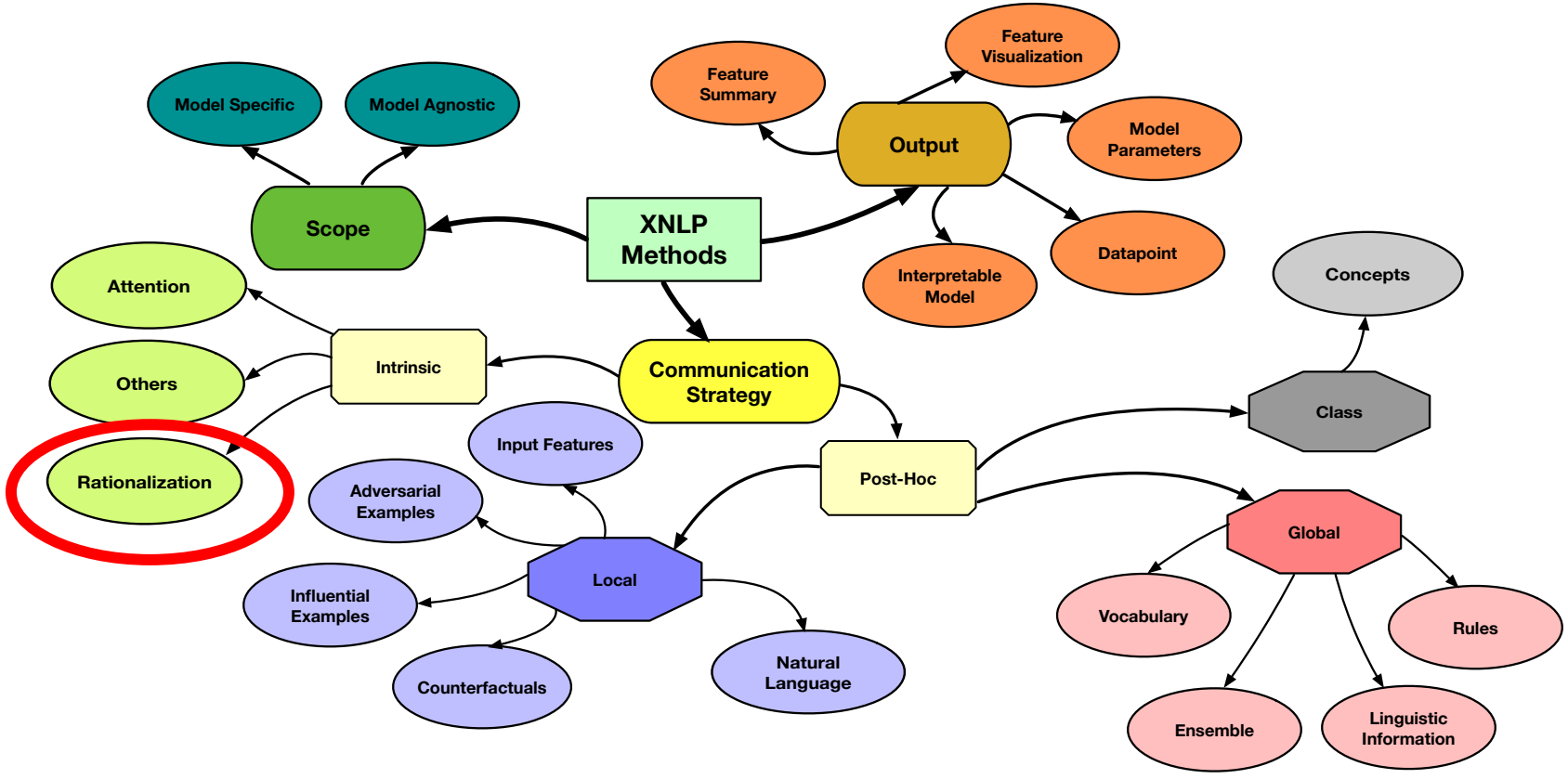


Compositional Explanations of Neurons, Mu and Andreas, 2020, http://arxiv.org/abs/2006.14032

XNLP Methods Perspectives



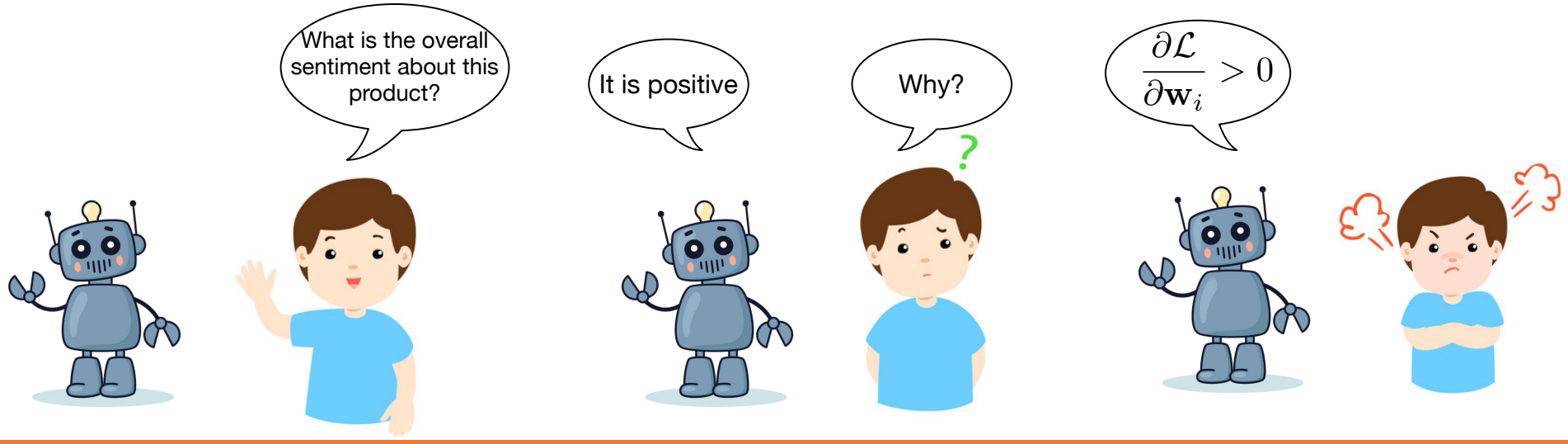
Rationalization Methods



Rationalizing Neural Predictions

Lei, et al., EMNLP 2016: https://aclanthology.org/D16-1011.pdf

- If a human takes a decision, they can easily explain their decision by providing justifications
- However, modern ML/NLP algorithms/models are excellent in making decisions but fail to explain in a human understandable way
- It would be nice of if the NLP model could explain itself by providing ***rationale*** for its decision
- Moreover, these rationale being in natural language can easily be understood by humans.



Images: Google Search

Generating Rationale

- Make rationale generation as an integral part of model prediction itself
- Extractive rationales: subset of tokens from the input
- Constraints:
  - Short and coherent pieces of text (e.g., phrases)
  - Rationales alone should be sufficient for prediction
- Unsupervised

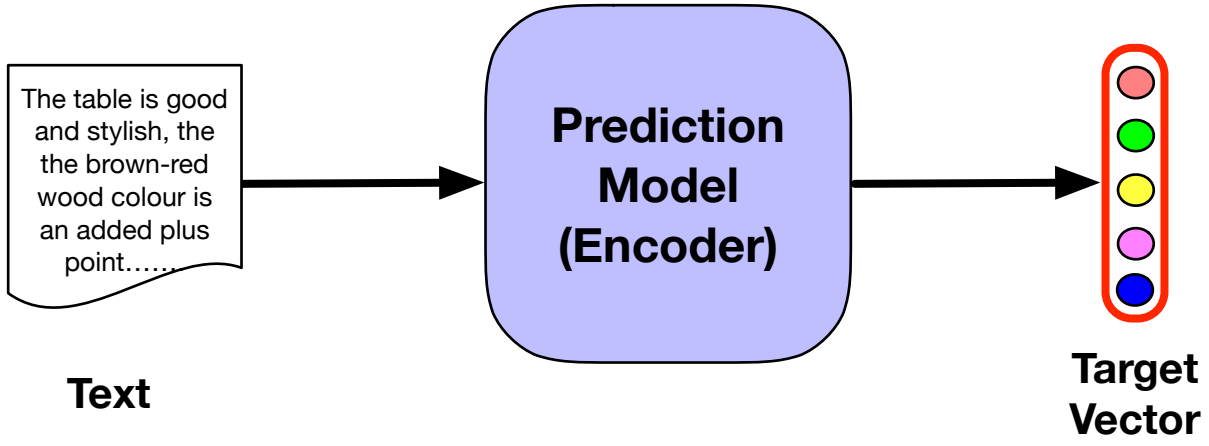
Review

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

Prediction Model

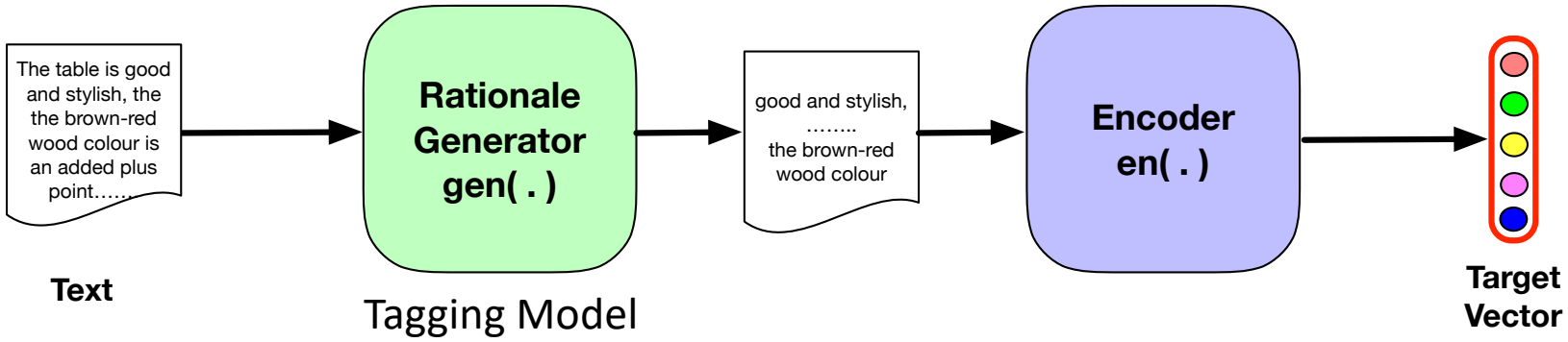


$$f : \mathbf{x} \rightarrow \mathbb{R}^m$$

$$\mathbf{x} = \{x_1, \ldots, x_l\} \qquad \mathbb{R}^m$$

$$x_i \in \mathbb{R}^d$$

Model for Generating Rationale



Encoder, enc(.)

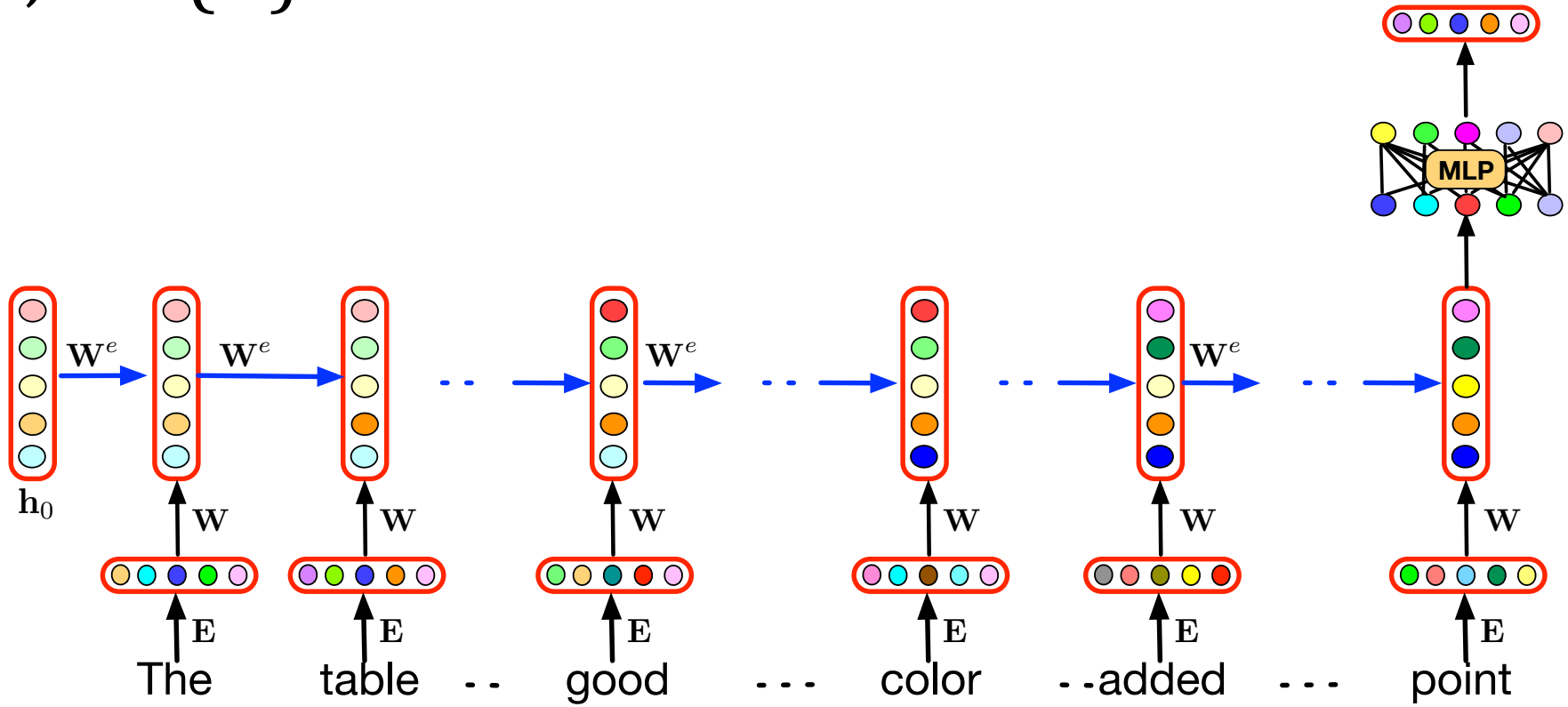
$$(\mathbf{x}, \mathbf{y})$$

$$\mathbf{x} = \{x_t\}_{t=1}^{l} \qquad \mathbf{y} \in [0,1]^m$$

$$\tilde{\mathbf{y}} = \mathbf{enc}(\mathbf{x})$$

$$\mathcal{L}(\mathbf{x}, \mathbf{y}) = || \tilde{\mathbf{y}} - \tilde{\mathbf{y}} ||_2^2 = || \mathbf{enc}(\mathbf{x}) - \tilde{\mathbf{y}} ||_2^2$$

Encoder, enc(.)



$$\mathbf{h}_t = f_e(\mathbf{x}_t, \mathbf{h}_{t-1})$$

$$\tilde{\mathbf{y}} = \sigma_e(\mathbf{W}^e \mathbf{h}_l + \mathbf{b}^e)$$

Generator, gen(.)

Binary Tagger: Tags each word with 0/1 depending on if it is part of rationale or not

$$\mathbf{x} = \{x_1, \ldots, x_l\} \longrightarrow \mathbf{z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_l\}$$

$$\mathbf{z}_t \in \{0, 1\}$$

$$\mathbf{z} \sim \mathbf{gen}(\mathbf{x}) \equiv p(\mathbf{z} \mid \mathbf{x})$$

Independence Assumption $$p(\mathbf{z} \mid \mathbf{x}) = \prod_{t=1}^{l} p(\mathbf{z}_t \mid \mathbf{x})$$

Generator, gen(.)

Components modelled via bi-directional RNNs

$$p(\mathbf{z} \mid \mathbf{x}) = \prod_{t=1}^{l} p(\mathbf{z}_t \mid \mathbf{x})$$

$$\overrightarrow{\mathbf{h}}_t = \overrightarrow{f}(x_t, \overrightarrow{\mathbf{h}}_{t-1})$$

$$\overleftarrow{\mathbf{h}}_t = \overleftarrow{f}(x_t, \overleftarrow{\mathbf{h}}_{t+1})$$

$$p(\mathbf{z}_t \mid \mathbf{x}) = \sigma_z(\mathbf{W}^{\mathbf{z}}[\overrightarrow{\mathbf{h}_t}; \overleftarrow{\mathbf{h}_t}] + \mathbf{b}^{\mathbf{z}})$$

Generator, gen(.)

Take into account the context

$$p(\mathbf{z} \mid \mathbf{x}) = \prod_{t=1}^{l} p(\mathbf{z}_t \mid \mathbf{x}, \mathbf{z}_1, \ldots, \mathbf{z}_{t-1})$$

$$p(\mathbf{z}_t \mid \mathbf{x}, \mathbf{z}_{1,t-1}) = \sigma_z(\mathbf{W}^{\mathbf{z}}[\overrightarrow{\mathbf{h}_t}; \overleftarrow{\mathbf{h}_t}; \mathbf{s}_{t-1}] + \mathbf{b}^{\mathbf{z}})$$

$$\mathbf{s}_t = f_z([\overrightarrow{h}; \overleftarrow{h}; \mathbf{z}_t], \mathbf{s}_{t-1})$$

Joint Objective Function

Generator and Encoder learned jointly

$$\tilde{y} = \mathbf{enc}(\mathbf{x}, \mathbf{z}) = \mathbf{enc}(\underbrace{\mathbf{gen}(x)}_{\mathbf{z}})$$

$$\mathcal{L}(\mathbf{z}, \mathbf{x}, \mathbf{y}) = || \ \mathbf{enc}(\mathbf{z}, \mathbf{x}) - \mathbf{y} \ ||_2^2$$

$$\Omega(\mathbf{z}) = \lambda_1 \ || \ \mathbf{z} \ || + \lambda_2 \sum_t | \ \mathbf{z}_t - \mathbf{z}_{t-1} \ |$$

$$\min_{\theta_e, \theta_g} \sum_{\mathbf{x}, \mathbf{y} \in D} \mathbb{E}_{\mathbf{z} \sim \mathbf{gen}(\mathbf{x})} [\underbrace{cost(\mathbf{z}, \mathbf{x}, \mathbf{y})}_{\mathcal{L}(\mathbf{z}, \mathbf{x}, \mathbf{y}) + \Omega(\mathbf{z})}]$$

Minimization of Joint Loss

$$\frac{\partial \mathbb{E}_{\mathbf{z} \sim \mathbf{gen}(\mathbf{x})} cost(\mathbf{z}, \mathbf{x}, \mathbf{y})}{\partial \theta_g} = \sum_{\mathbf{z}} cost(\mathbf{z}, \mathbf{x}, \mathbf{y}) . \frac{\partial p(\mathbf{z} \mid \mathbf{x})}{\partial \theta_g}$$

$$= \sum_{\mathbf{z}} cost(\mathbf{z}, \mathbf{x}, \mathbf{y}) . \frac{\partial p(\mathbf{z} \mid \mathbf{x})}{\partial \theta_g} . \frac{p(\mathbf{z} \mid \mathbf{x})}{p(\mathbf{z} \mid \mathbf{x})}$$

$$= \sum_{\mathbf{z}} cost(\mathbf{z}, \mathbf{x}, \mathbf{y}) \frac{\partial log\ p(\mathbf{z} \mid \mathbf{x})}{\partial \theta_g} . p(\mathbf{z} \mid \mathbf{x})$$

$$= \mathbb{E}_{\mathbf{z} \sim \mathbf{gen}(\mathbf{x})} \left[cost(\mathbf{z}, \mathbf{x}, \mathbf{y}) \frac{\partial log\ p(\mathbf{z} \mid \mathbf{x})}{\partial \theta_g} \right]$$

Minimization of Joint Loss

$$\min_{\theta_e,\theta_g} \sum_{\mathbf{x},\mathbf{y}\in D} \mathbb{E}_{\mathbf{z}\sim\mathbf{gen}(\mathbf{x})}[\underbrace{cost(\mathbf{z},\mathbf{x},\mathbf{y})}_{\mathcal{L}(\mathbf{z},\mathbf{x},\mathbf{y})+\Omega(\mathbf{z})}]$$

$$\frac{\partial \mathbb{E}_{\mathbf{z}\sim\mathbf{gen}(\mathbf{x})} cost(\mathbf{z},\mathbf{x},\mathbf{y})}{\partial \theta_g} = \mathbb{E}_{\mathbf{z}\sim\mathbf{gen}(\mathbf{x})}\left[cost(\mathbf{z},\mathbf{x},\mathbf{y})\frac{\partial log\ p(\mathbf{z}\mid\mathbf{x})}{\partial \theta_g}\right]$$

$$\frac{\partial \mathbb{E}_{\mathbf{z}\sim\mathbf{gen}(\mathbf{x})} cost(\mathbf{z},\mathbf{x},\mathbf{y})}{\partial \theta_e} = \mathbb{E}_{\mathbf{z}\sim\mathbf{gen}(\mathbf{x})}\left[\frac{\partial cost(\mathbf{z},\mathbf{x},\mathbf{y})}{\partial \theta_g}\right]$$

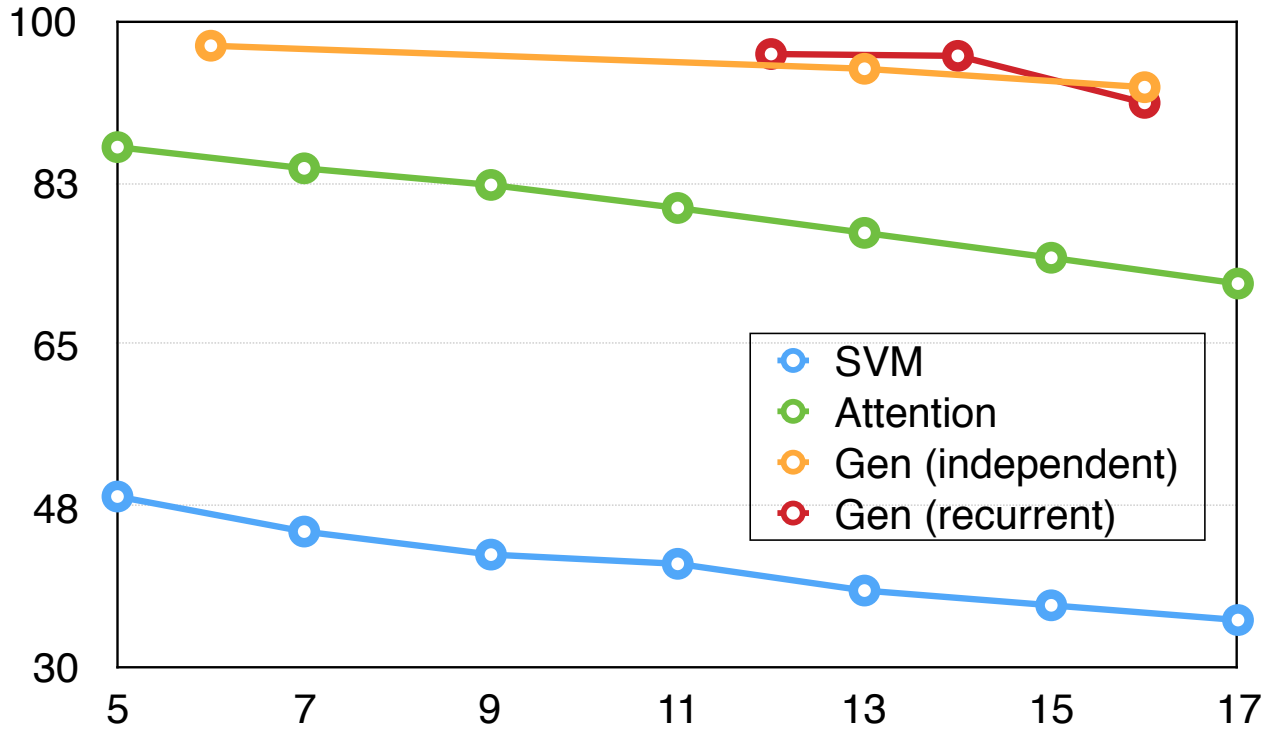
Experiments: Multi-aspect Sentiment Analysis

BeerAdvocate Review Dataset

Aspects: Appearance, Aroma, Palate, Taste

Rating (1 to 5) for each aspect and overall rating

Experiments: Multi-aspect Sentiment Analysis



Precision vs Percentage of Text as Rationales for Appearance aspect

Lei, et al., EMNLP 2016

Experiments: Multi-aspect Sentiment Analysis

a beer that is not sold in my neck of the woods , but managed to get while on a roadtrip . poured into an imperial pint glass with a generous head that sustained life throughout . nothing out of the ordinary here , but a good brew still . body was kind of heavy , but not thick . the hop smell was excellent and enticing . very drinkable

---

very dark beer . pours a nice finger and a half of creamy foam and stays throughout the beer . smells of coffee and roasted malt . has a major coffee-like taste with hints of chocolate . if you like black coffee , you will love this porter . creamy smooth mouthfeel and definitely gets smoother on the palate once it warms . it 's an ok porter but i feel there are much better one 's out there .

---

i really did not like this . it just seemed extremely watery . i dont ' think this had any carbonation whatsoever . maybe it was flat , who knows ? but even if i got a bad brew i do n't see how this would possibly be something i 'd get time and time again . i could taste the hops towards the middle , but the beer got pretty nasty towards the bottom . i would never drink this again , unless it was free . i 'm kind of upset i bought this .

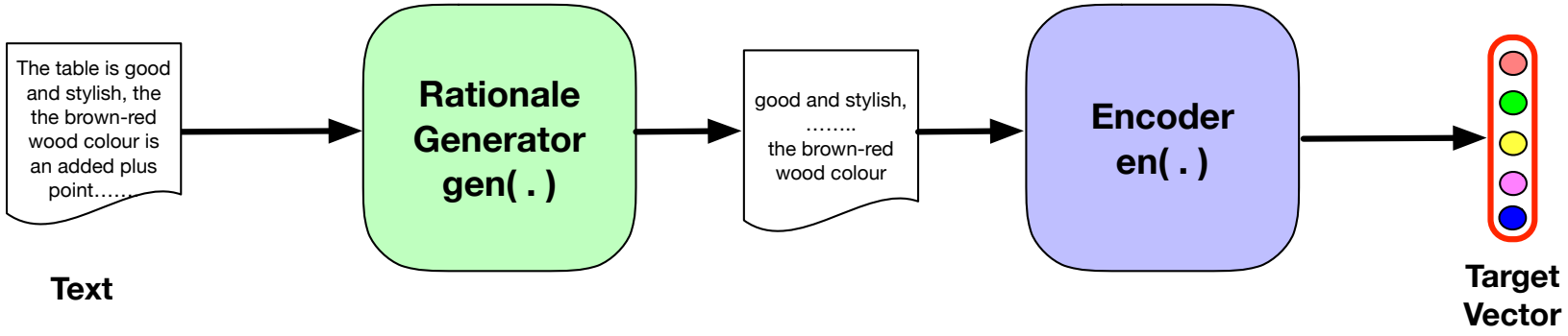
---

a : poured a nice dark brown with a tan colored head about half an inch thick , nice red/garnet accents when held to the light . little clumps of lacing all around the glass , not too shabby . not terribly impressive though s : smells like a more guinness-y guinness really , there are some roasted malts there , signature guinness smells , less burnt though , a little bit of chocolate … … m : relatively thick , it is n't an export stout or imperial stout , but still is pretty hefty in the mouth , very smooth , not much carbonation . not too shabby d : not quite as drinkable as the draught , but still not too bad . i could easily see drinking a few of these .

Lei, et al., EMNLP 2016

Summary

- One of the first method to propose generating natural language rationales easily understandable by humans.
- Explain and then Predict: Fairly modular and generic framework.



Explain and Predict, and then Predict Again

Zhang, et al., WSDM 2021: https://arxiv.org/abs/2101.04109

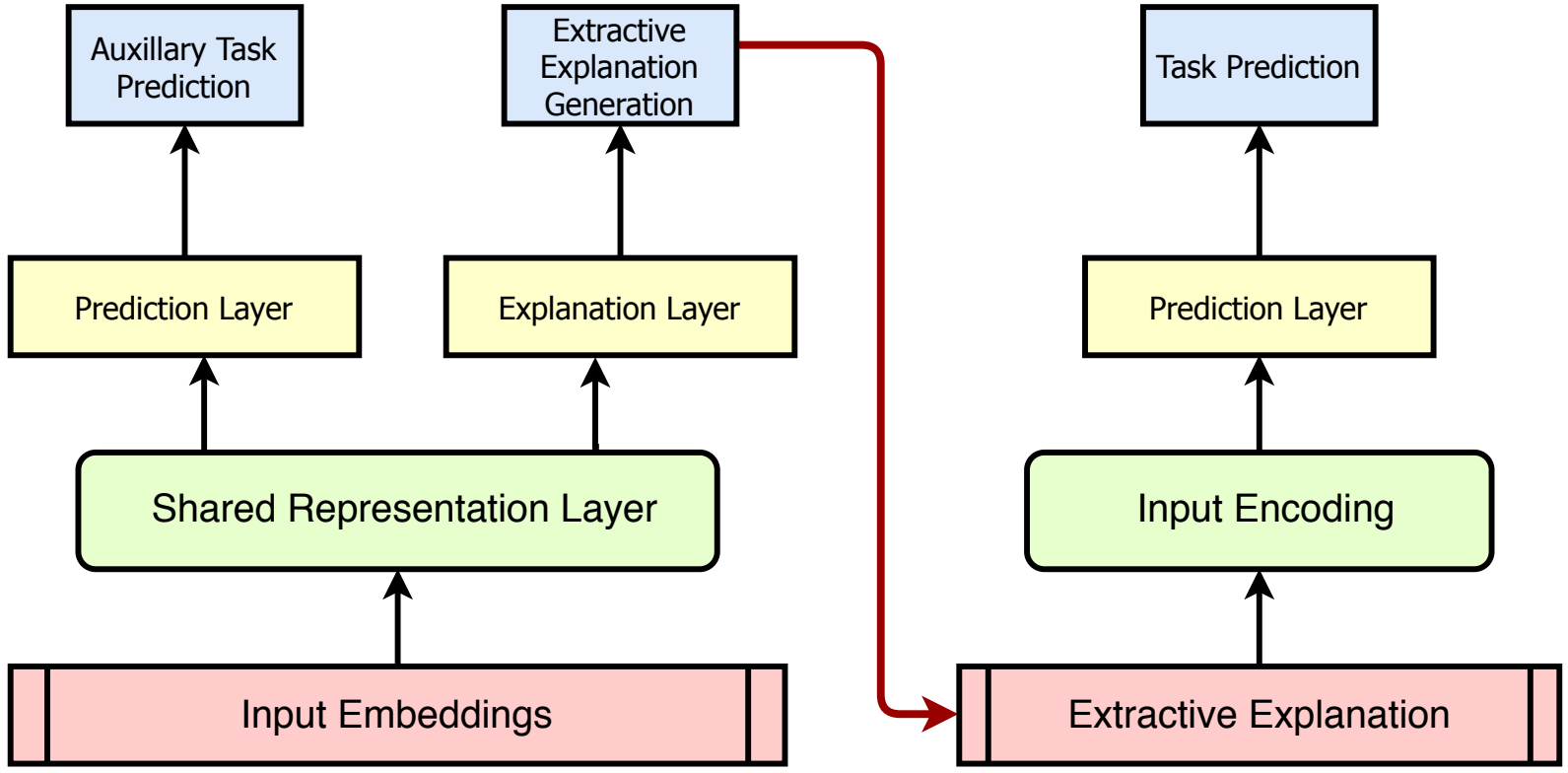
- Rationale based methods first generate explanation and then predict
- These methods only use input text for supervision
- What if, for an application, you had access to rationales data (e.g., FEVER dataset). This could provide additional inductive bias to improve task performance.
- **ExPred**: Multi-task learning during the explanation generation phase, trading off explanation and prediction losses
- Interpretable by Design

Query: **san francisco bay area contains zero towns**

Label: REFUTE, **Predict:** REFUTE

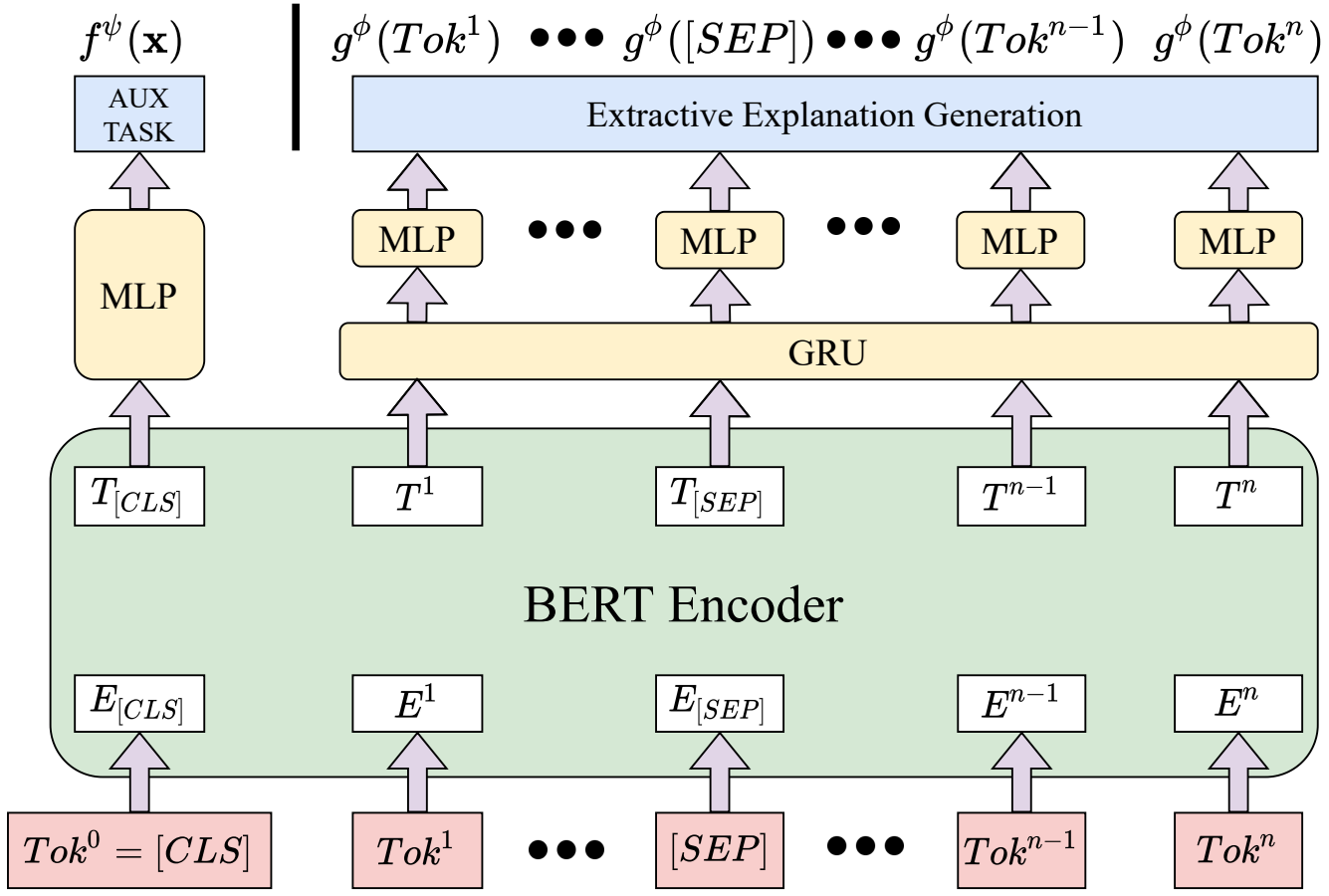
Input Passage: the san francisco bay area, referred to locally as the bay area is a populous region surrounding the san francisco and san pablo estuaries in northern california. The region encompasses the major cities and metropolitan areas of san jose, san francisco, and Oakland, along with smaller urban and rural areas. The bay area's nine counties are .....Santa Clara, Solana and Sonoma. Home to approximately 7.68 million people, the nine-county bay area contained many cities, towns, airports, and associated regional, state, and national parks, connected by a network of roads, highways, railroads, bridges, tunnels and commuter rail. The combined statistical area of the region is the second largest in california after the Los Angeles area.

ExPred



Zhang, et al., WSDM 2021

Explanation Generation



Zhang, et al., WSDM 2021

Loss Function

$$\mathcal{L}_{exp} = \frac{1}{\mid S \mid} \sum_{i=1}^{|S|} \mid S_{t^i} \mid . \, BCE(p^i, t^i)$$

$$\mathcal{L}_{loss} = \mathcal{L}_{task} + \lambda \mathcal{L}_{exp}$$

Prediction Task

$$f^{\psi}(\mathbf{x}) \rightarrow \{0,1\}$$

$$g^{\phi}(\mathbf{x}) \rightarrow \{0,1\}^{|S|}$$

$$f^{\theta}(\mathbf{x}\ g^{\phi}(\mathbf{x})) \rightarrow \{0,1\} \quad if\ f\ f^{\psi}(x) = y$$

Experiments: FEVER (Fact Extraction and VERification)

- Fact Checking Dataset
- The task is to verify claims from textual sources
- Each claim is to be classified as *Supported, Refuted*, or *Not Enough Info*
- Human annotated explanations

Experiments: FEVER (Fact Extraction and VERification)

Approaches	FEVER	
	Macro F1	**Token F1**
DeYoung et al. [7]	0.719	0.234
Lei et al. [19]	0.718	-[1]
Lehman et al. [18]	0.691	0.523
Bert-To-Bert	0.877	0.812
EXPRED-STAGE-1	**0.907**	**0.837**
EXPRED (W/O TASK SUP.)	0.795	0.801
EXPRED	0.894	**0.837**
HUMAN EXPLANATION	0.921	1.0
FULL INPUT	0.916	-

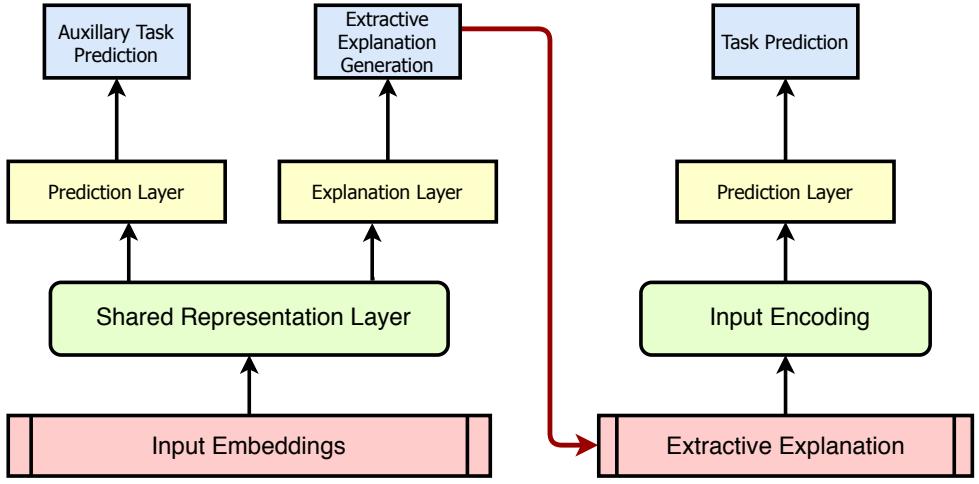
Experiments: FEVER (Fact Extraction and VERification)

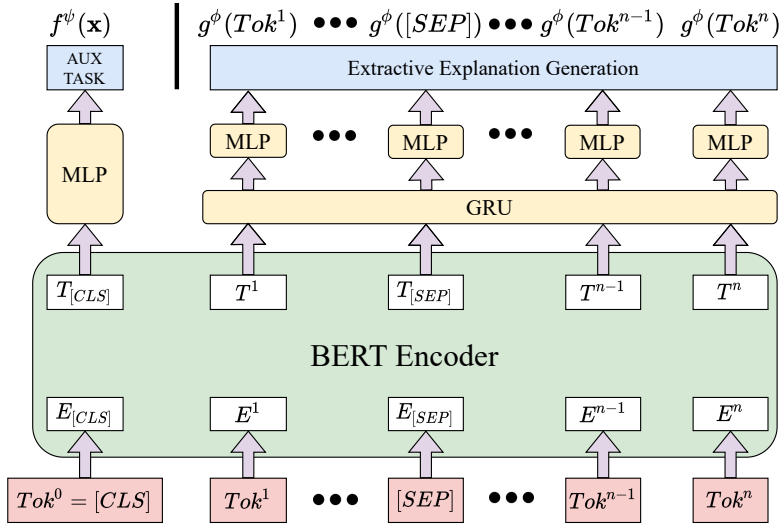
Claim: Emma Watson was killed in 1990.

Human	Evidence: Emma Charlotte Duerre Watson (born 15 april 1990) is a French-British actress, model, and activist. Born in Paris ... previously. Watson appeared in all eight Harry Potter films from 2001 to 2011, earning worldwide fame, critical accolades, and around $60 million..
Expred(w/o task)	Evidence: Emma Charlotte Duerre Watson (born 15 april 1990) is a French-British actress, model, and activist. Born in Paris ... previously. Watson appeared in all eight Harry Potter films from 2001 to 2011, earning worldwide fame, critical accolades, and around $60 million..
Expred	Evidence: Emma Charlotte Duerre Watson (born 15 april 1990) is a French-British actress, model, and activist. Born in Paris ... previously. Watson appeared in all eight Harry Potter films from 2001 to 2011, earning worldwide fame, critical accolades, and around $60 million..
Lehman et al.	Evidence: Emma Charlotte Duerre Watson (born 15 april 1990) is a French-British actress, model, and activist. Born in Paris ... previously. Watson appeared in all eight Harry Potter films from 2001 to 2011, earning worldwide fame, critical accolades, and around $60 million..
Bert to Bert	Evidence: Emma Charlotte Duerre Watson (born 15 april 1990) is a French-British actress, model, and activist. Born in Paris ... previously. Watson appeared in all eight Harry Potter films from 2001 to 2011, earning worldwide fame, critical accolades, and around $60 million..

Summary

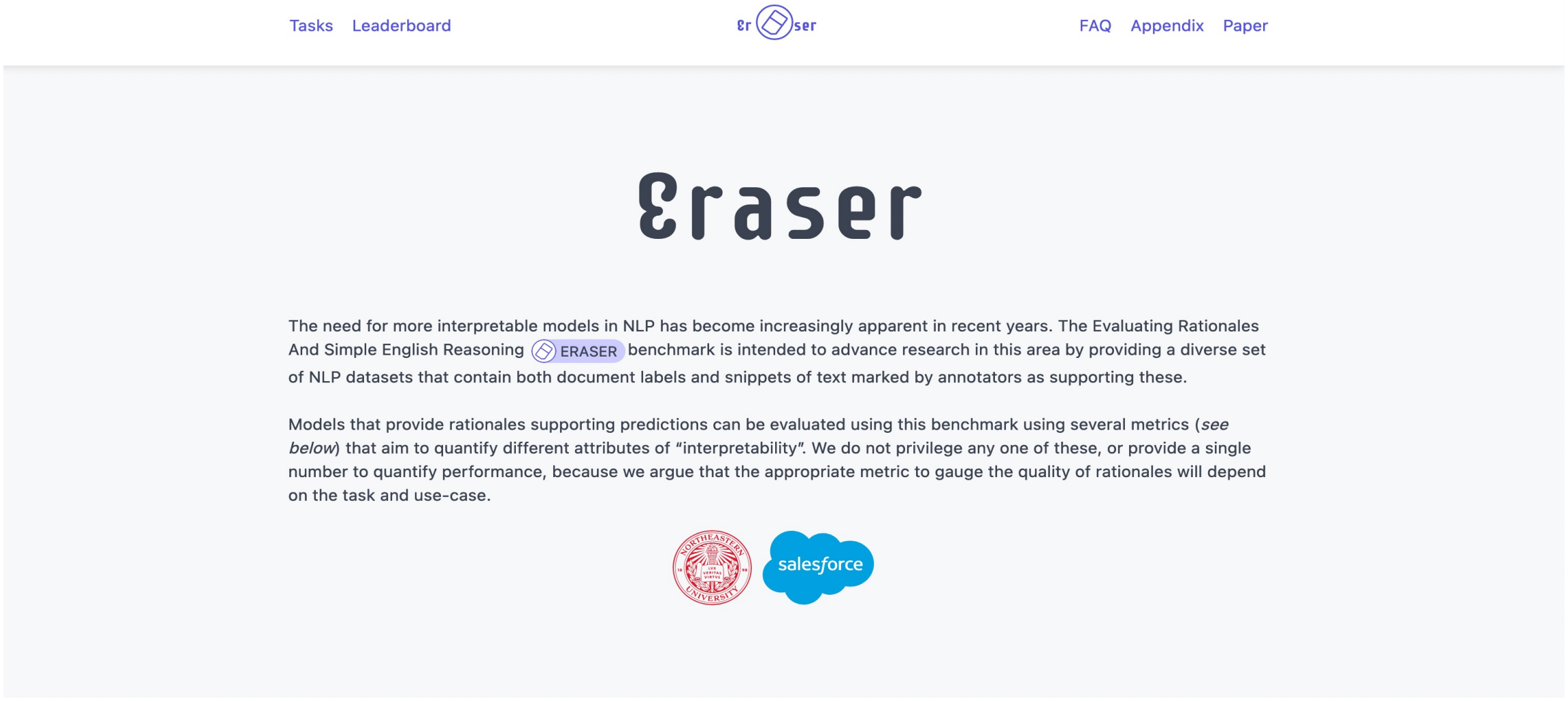
- Explanations should be *first-class-citizens*
- For an application if you have human explanation annotation, use them via MTL



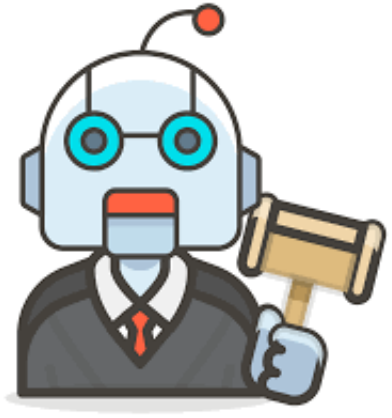


ERASER (Evaluating Rationales And Simple English Reasoning) Benchmark

(https://www.eraserbenchmark.com/)



Legal-NLP



Why do Legal Texts need Special Tools?

Why do Legal Texts need Special Tools?

- Long Documents

Why do Legal Texts need Special Tools?

- Long Documents
- Domain Specific Lexicon

Legal Text vs Non-Legal Text

Domain Specific Lexicon

4. As the factual narration would evince, on 10th February, 2016, a team of assessors of the respondent No. 2 conducted verification assessment for grant of LOP for the academic year 2016-17. In the mean time, the Constitution Bench in Modern Dental College and Research Center and others v. State of Madhya Pradesh and others 1 constituted the Oversight Committee headed by Justice R.M. Lodha former CJI to oversee the functioning of the MCI. We shall refer the relevant paragraphs of the said judgment at a later stage. On 13th May, 2016, the report of the assessors team was considered by the Executive Committee of the respondent No.2 in its meeting dated 13.05.2016 and on 14.5.2016 the MCI recommended the disapproval of the scheme of the petitioner under Section 10-A of the Act for the academic year 2016-17. However, after Oversight Committee was constituted, the Central Government issued a public notice informing all the Medical 1 (2016) 7 SCC 353 Colleges to submit a compliance report concerning their respective colleges who had applied for LOP for 2016-17. As the facts would unfold, the 1st respondent sent the compliance report along with the reply of the MCI to the Oversight Committee for consideration which on 11.08.2016 approved the same for the year 2016-17 imposing certain conditions.

Legal and Formal text terminology

Why do Legal Texts need Special Tools?

- Long Documents
- Domain Specific Lexicon
- Highly unstructured and noisy

Legal Text vs Non-Legal Text

Noisy and unstructured text

20. Section 10-A of the Act deals with permission for establishment of new medical college, new course of study, etc. Sub-section (7) of Section 10-A reads as follows:-

"(7) The Council, while making its recommendations under clause (b) of sub-

section (3) and the Central Government, while passing an order, either approving or disapproving the scheme under sub-section (4), shall have due regard to the following factors, namely—

(a) whether the proposed medical college or the existing medical college seeking to open a new or higher course of study or training, would be in a position to offer the minimum standards of medical education as prescribed by the Council under Section 19A or, as the case may be, under Section 20 in the case of postgraduate medical education.

Document contains some relevant information in the main text.

Why do Legal Texts need Special Tools?

- Long Documents
- Domain Specific Lexicon
- Highly unstructured and noisy
- Legal domain is further divided into non-overlapping sub-domains

Why do Legal Texts need Special Tools?

- Long Documents
- Domain Specific Lexicon
- Highly unstructured and noisy
- Legal domain is further divided into non-overlapping sub-domains
- **Models need to be explainable**

Why do Legal Texts need Special Tools?

- Long Documents
- Domain Specific Lexicon
- Highly unstructured and noisy
- Legal domain is further divided into non-overlapping sub-domains
- **Models need to be explainable**
- Law is country/region specific

What is lacking?

- Specialized corpora
- Pre-trained models
- Basic tools to process texts

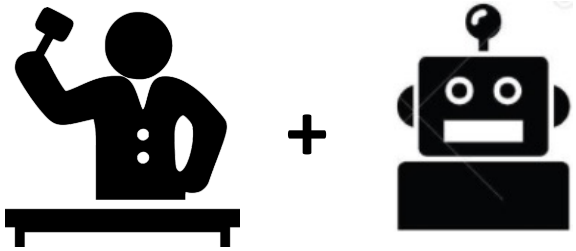
Typical Applications/Tasks in Legal Domain

- Prior Case Retrieval
- Summarization
- Argumentation Mining
- Information Extraction and Retrieval
- Semantic Segmentation of documents
- Legal Judgement Prediction

ILDC for CJPE:Indian Legal Documents Corpus for Court Judgment Prediction and Explanation

Malik, et al., ACL 2021: https://aclanthology.org/2021.acl-long.313/

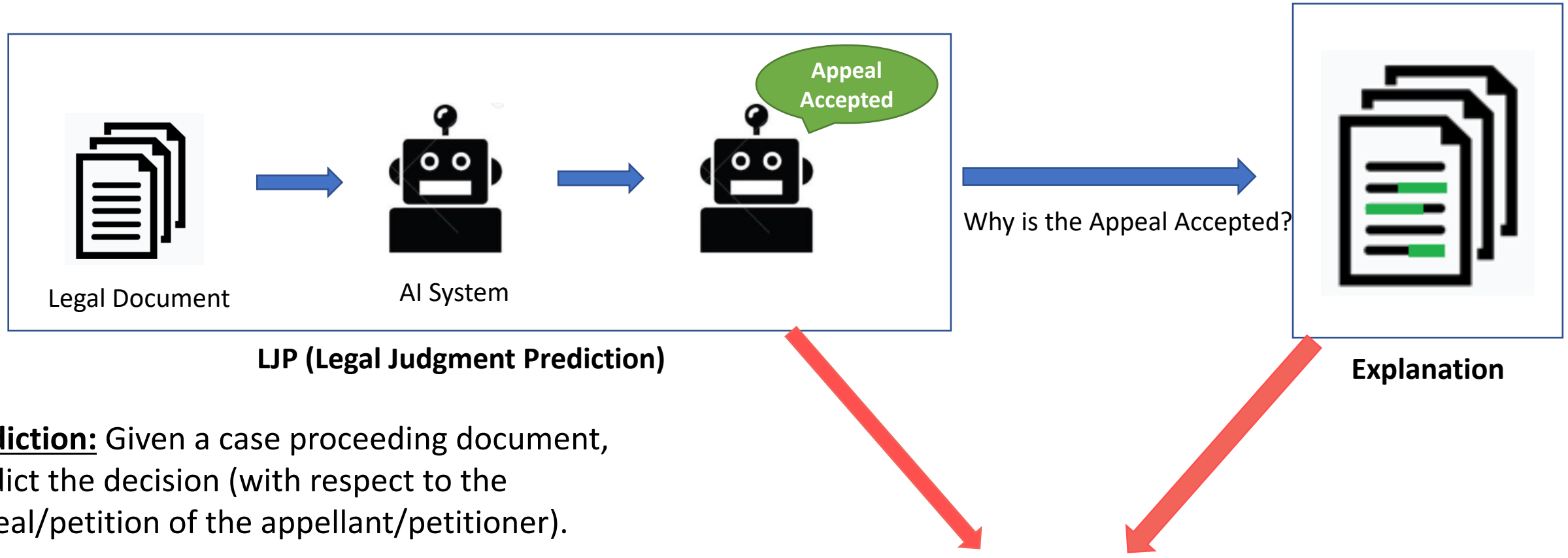
- In many highly populated countries (e.g., India), there exist a vast number of **pending backlog of legal cases** (Katju, 2019) that impede the judicial process.



- A system capable of **assisting/augmenting a judge** by suggesting the outcome of the court case will be useful.

- For the system to be of **practical utility**, in addition to prediction, it should **explain that outcome** in terms of how a legal practitioner understands the legal process.

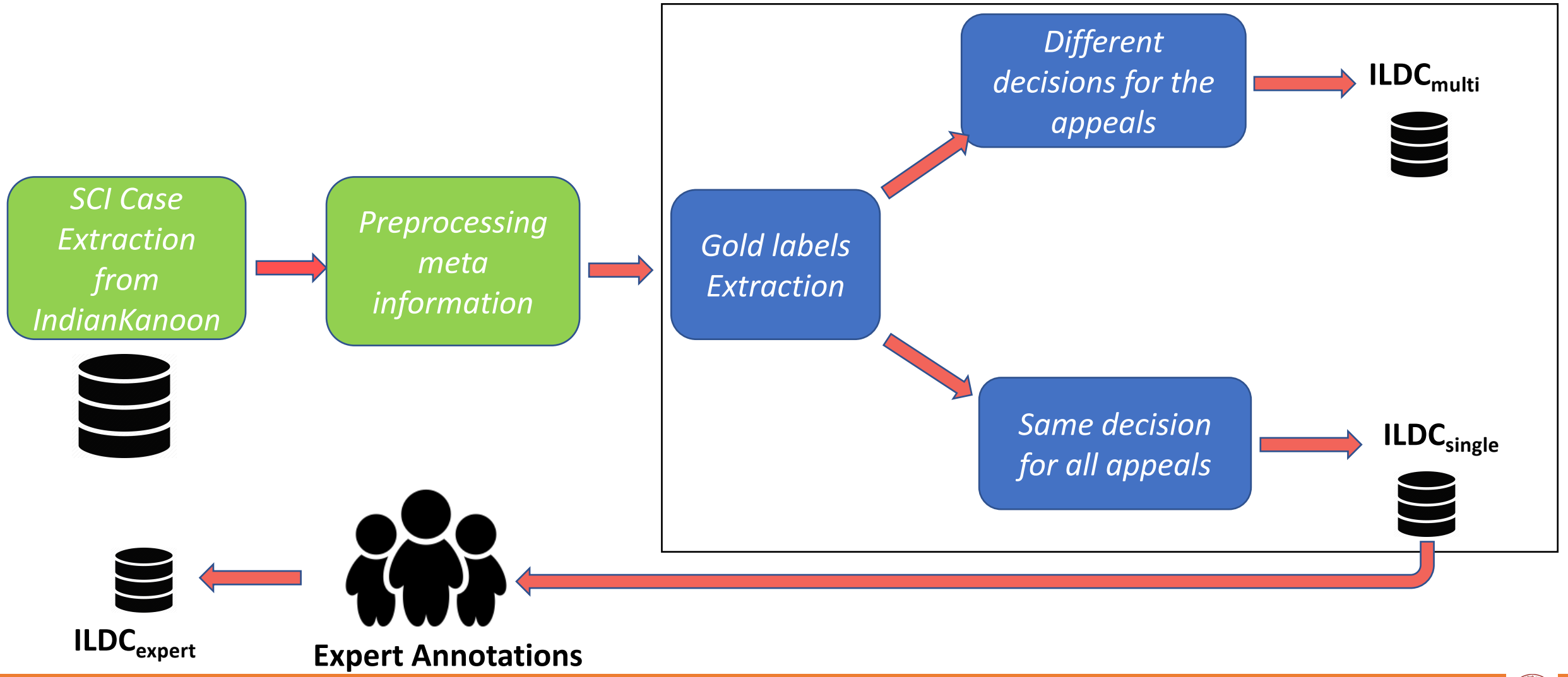
CJPE Task



Prediction: Given a case proceeding document, predict the decision (with respect to the appeal/petition of the appellant/petitioner).

Explanation: For predicted decision for the case, explain the decision by predicting important sentences that lead to the decision.

ILDC Creation



ILDC Statistics

Corpus	Average tokens	Number of documents (%Accepted cases)		
		Train	Validation	Test
$ILDC_{multi}$	3231	32305 (41.43%)	994 (50%)	1517 (50.23%)
$ILDC_{single}$	3884	5082 (38.08%)		
$ILDC_{expert}$	2894	56 (51.78%)		

Temporal Aspects and Bias handling in ILDC

- Train-Test-Validation split was *NOT* based on any temporal consideration: (Why?)
  - **System's aim** is to **identify standard features** of judgments
  - **If not**, such a system is likely to **fail** in application since, in the future **laws might get amended or replaced.**
- We have not made any specific choice about any **specific law** or any **category of cases** (random sampling).
- Names of the **judge(s), appellants, petitioners**, etc., were **anonymized**. (**Why?**):
  - A **strong indicator** of the case outcome
  - The system should focus on the **facts and applicable law.**

Annotation Analysis for Judgment Prediction

Quantitative Analysis

- Pairwise inter-expert agreement for judgment prediction were as **low as 85.7%** and as **high as 94.6%.**
- Fleiss kappa between all 5 annotators: **0.82**

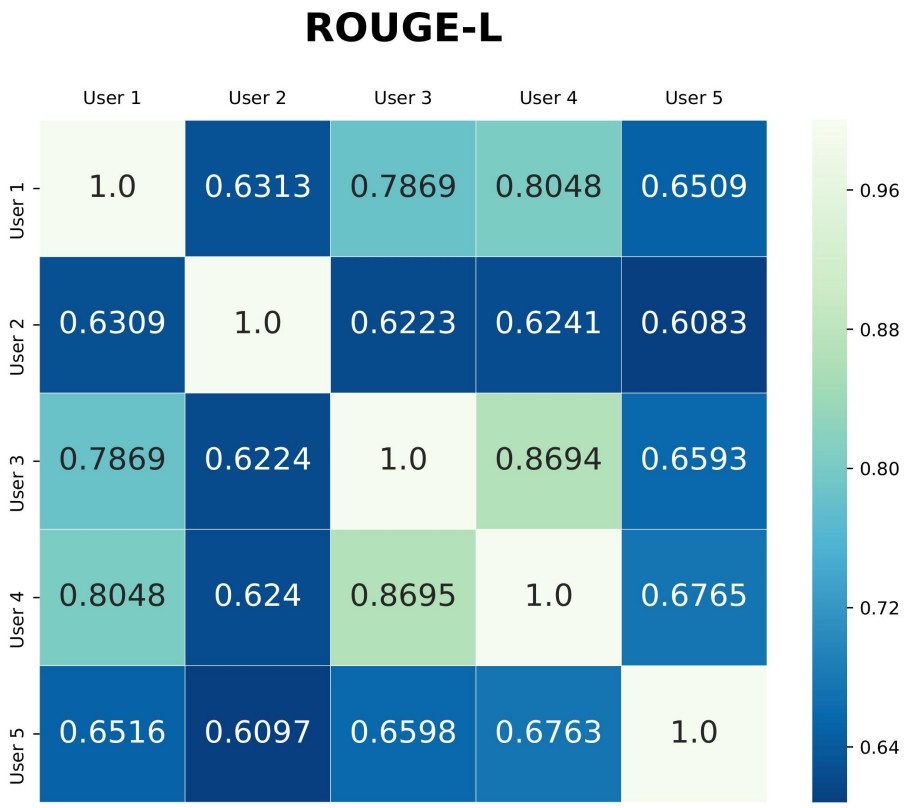
Expert	Accuracy(%)
Expert 1	94.64
Expert 2	91.07
Expert 3	98.21
Expert 4	89.28
Expert 5	96.42
Average	93.92

Qualitative Analysis

Our case studies point out: sources of confusion are mainly due to differences in linguistic interpretation (by the experts) of the legal language given in the case document.

Annotation Analysis for Explanations

Quantitative Analysis



Qualitative Analysis

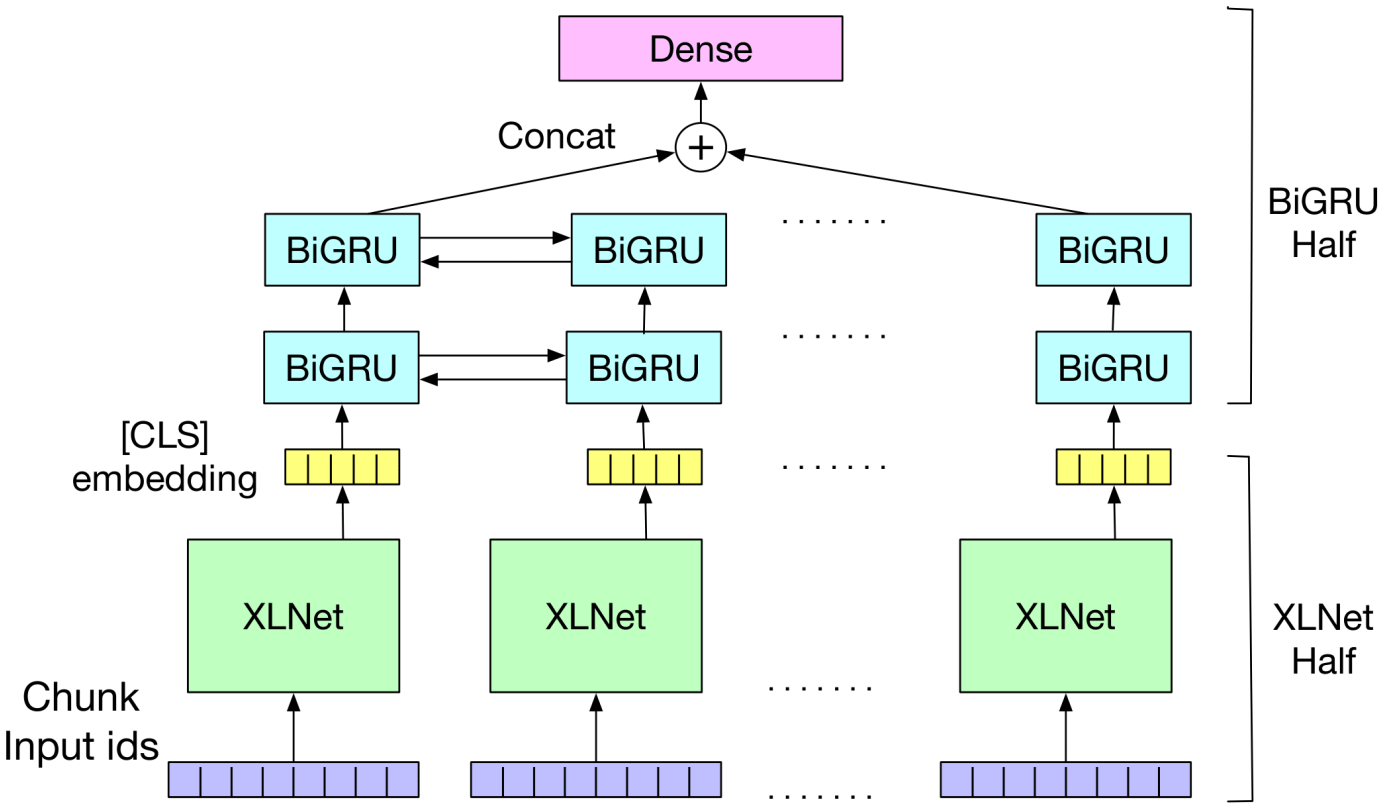
- Expert 2 and Expert 5 use **bare-minimum reasoning**
- Expert 1, Expert 3, and Expert 4 consider **holistic reasoning** for the decision. (**Substantive + Procedural aspects)**

Judgement Prediction

Model	Macro F1(%)	Accuracy(%)
Sequential Models		
Sent2vec + BiGRU + attention	59.66	58.31
GLoVe + BiGRU + attention	**64.35**	60.75
HAN	59.77	59.53
Transformers Models		
BERT Base (first 512 tokens)	59.06	57.65
BERT Base (last 512 tokens)	68.31	67.24
RoBERTa	**71.77**	71.26
XLNet	71.07	70.01
Hierarchical Models		
XLNet + BiGRU	**77.79**	**77.78**
Hierarchical Models with Attention		
XLNet + BiGRU + attention	**77.07**	77.01

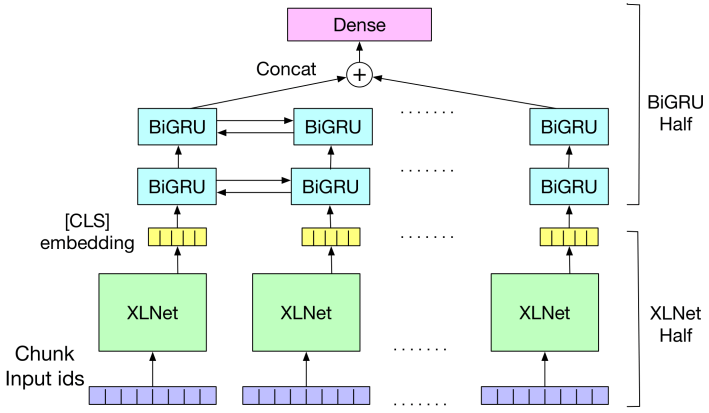
Annotator	Accuracy (%)
Average	93.92

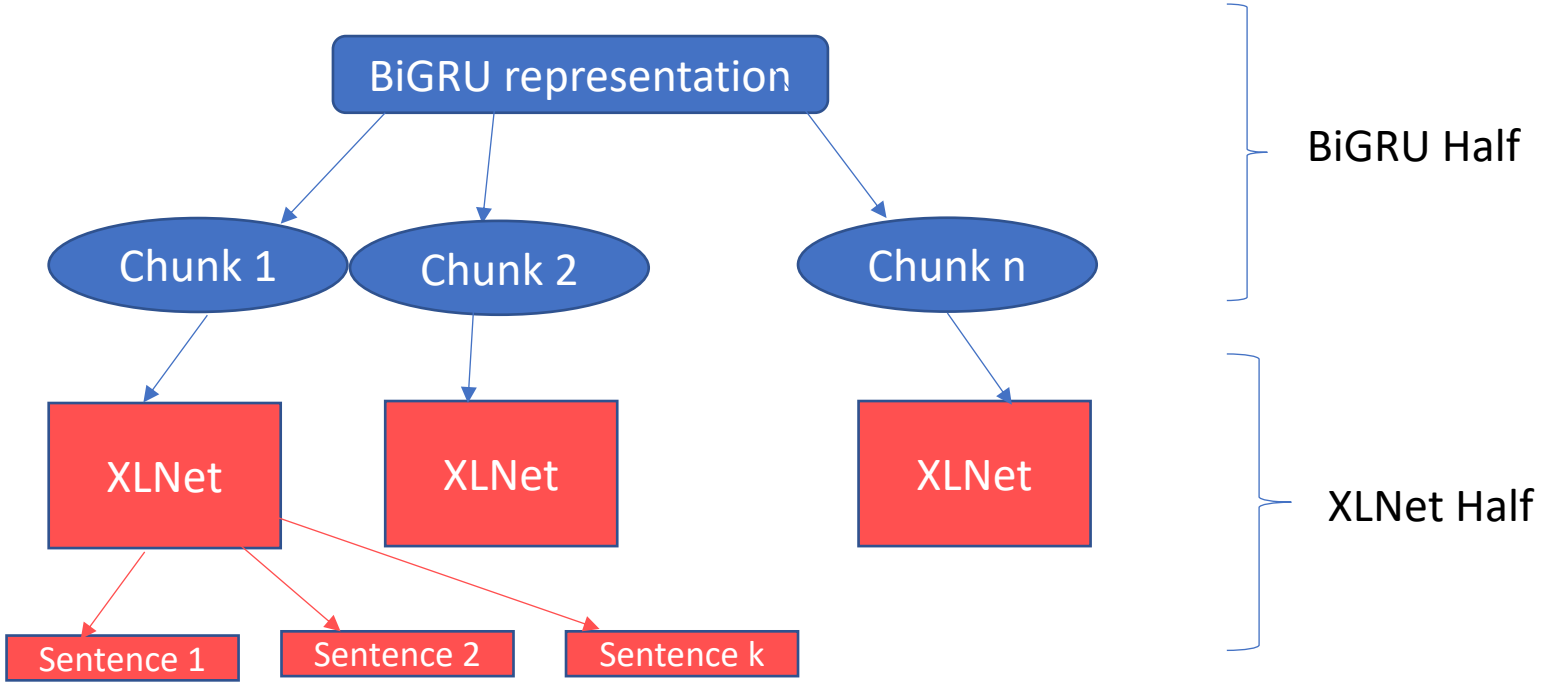
Prediction Model



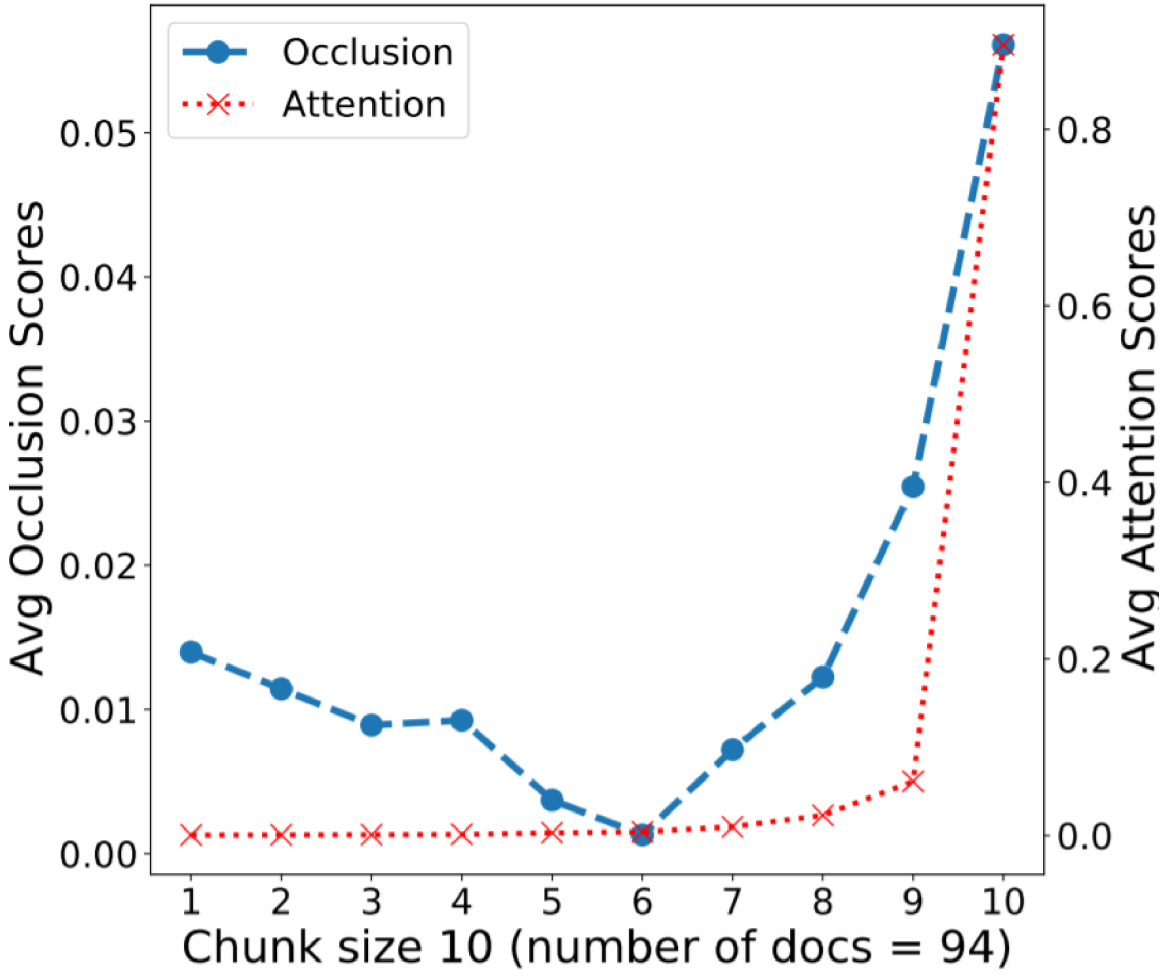
Explanation Methodology

- **Mask** a chunk embedding
- Calculate **masked probability of the label**
- **Compare** with **unmasked probability** and get the chunk importance score.



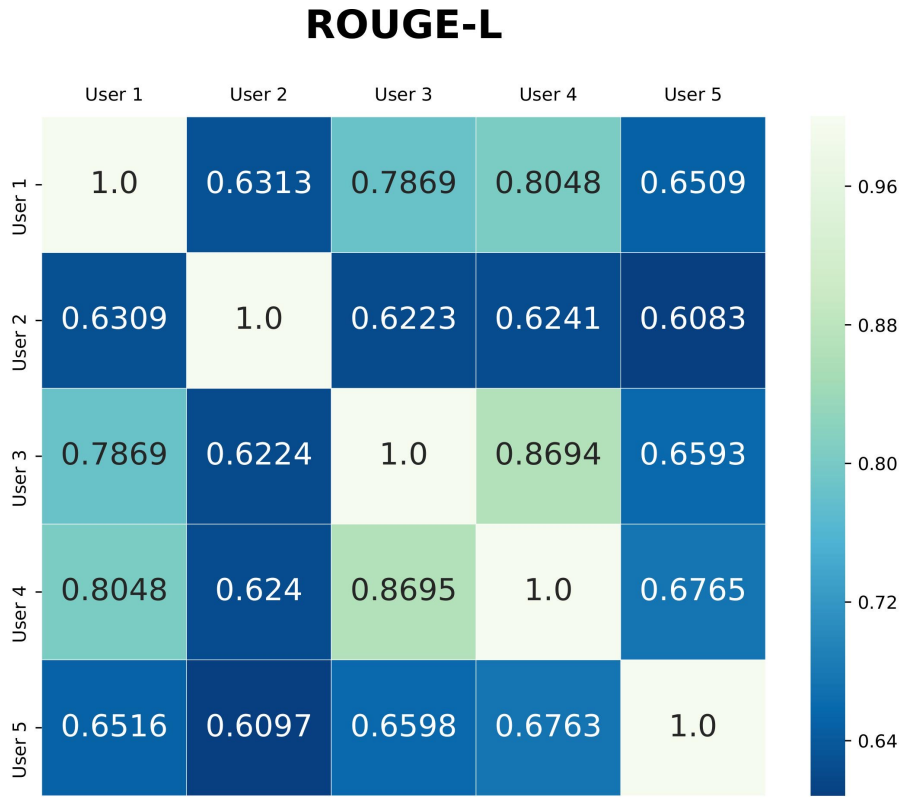


Insights from chunk scores



Judgment explanation

Metric	Explainability Model vs Experts				
	Expert				
	1	2	3	4	5
Jaccard Similarity	0.333	0.317	0.328	0.324	0.318
Overlap-Min	0.744	0.589	0.81	0.834	0.617
ROUGE-L	0.439	0.407	0.423	0.444	0.407
BLEU	0.16	0.28	0.099	0.093	0.248
Meteor	0.22	0.3	0.18	0.177	0.279



Summary

- Legal NLP is a new and upcoming area. Lot of work needs to be done from grounds up.
- This talk introduced the **ILDC corpus** and the **CJPE task**
- The corpus is **annotated** with decisions and their explanations for a separate test set (**$ILDC_{expert}$**)
- Analysis of the corpus and modeling results **shows the complexity of legal documents**
- Code Repository: https://github.com/Exploration-Lab/CJPE
- Hands-on Session

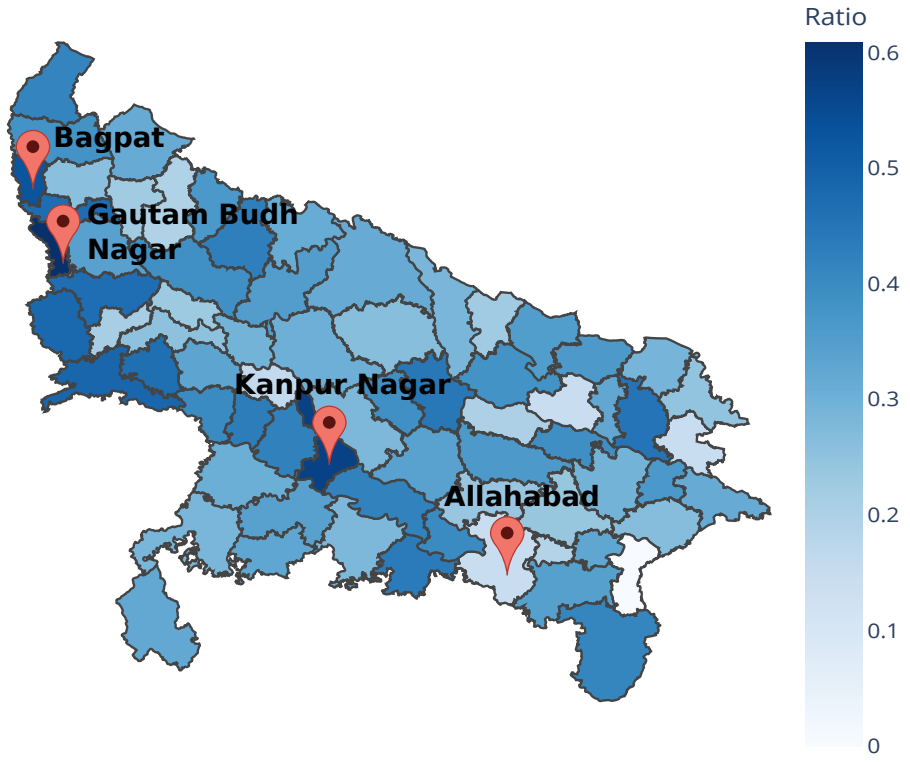
HLDC: Hindi Legal Document Corpus

Kapoor, et al., ACL 2021: https://aclanthology.org/2022.findings-acl.278/

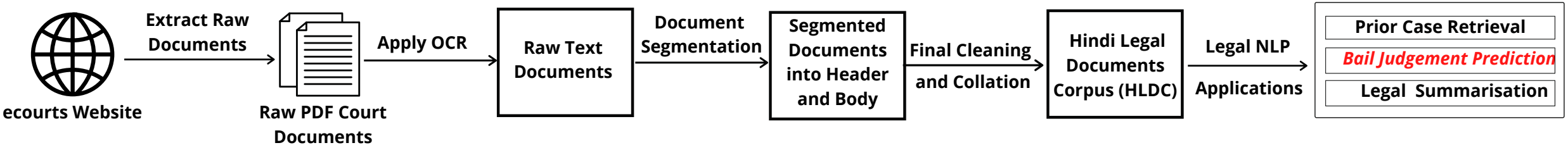
- Hindi: Language spoken by approx. 567 million people
- Legal NLP models developed on English do not transfer well to Hindi
- Corpus of 900K+ documents in Hindi
- ***Bail Prediction Task***: Given the facts of the case, predict if the bail should be granted or denied.

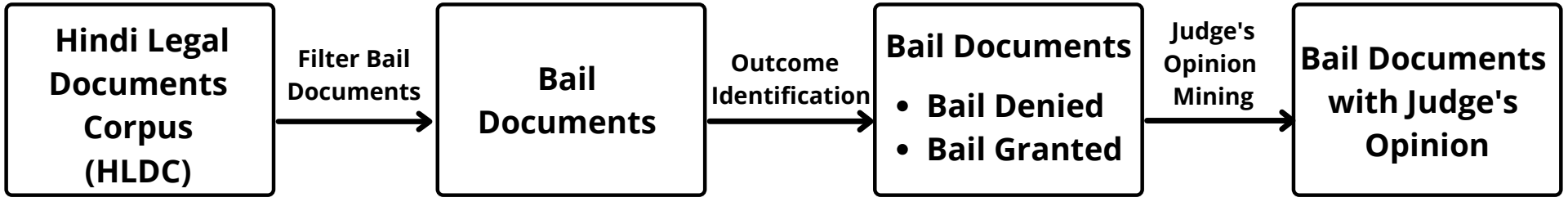
Why Bail Prediction?

Case Type	% in HLDC
Bail Applications	31.71
Criminal Cases	20.41
Original Suits	6.54
Warrant or Summons in Criminal Cases	5.24
Complaint Cases	4.37
Civil Misc	3.4
Final Report	3.32
Civil Cases	3.23
Others (Matrimonial Cases, Session Trial, Motor Vehicle Act, etc.)	21.75

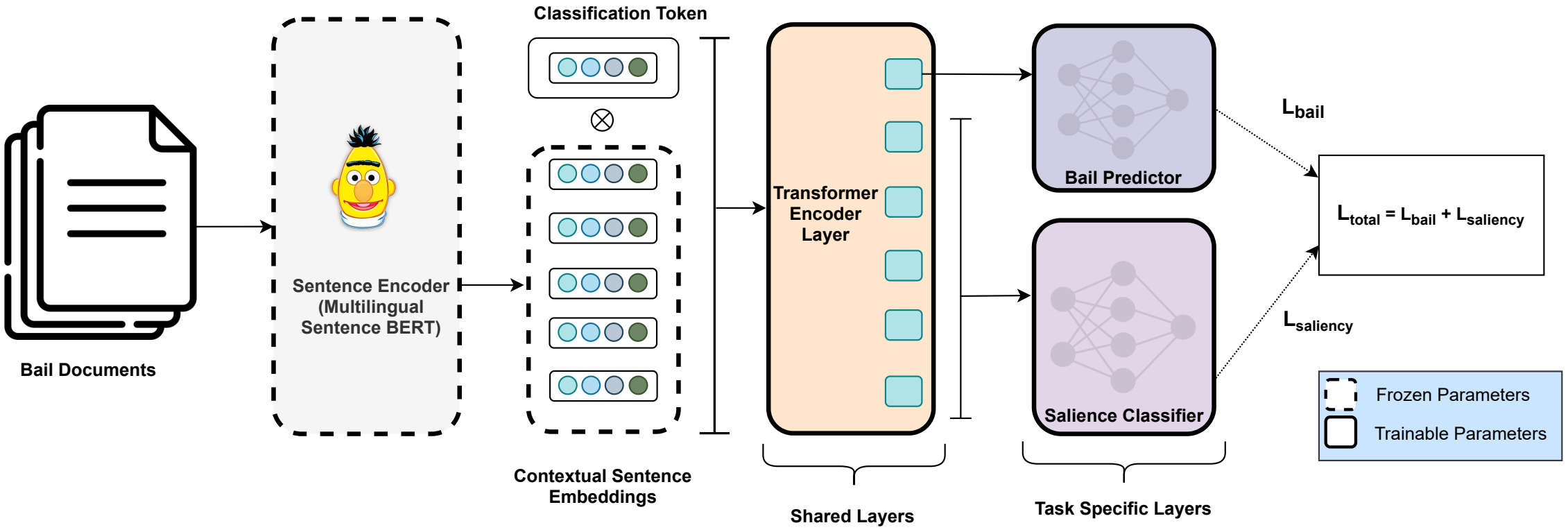


Corpus Creation





Multitask Bail Prediction Pipeline



Results

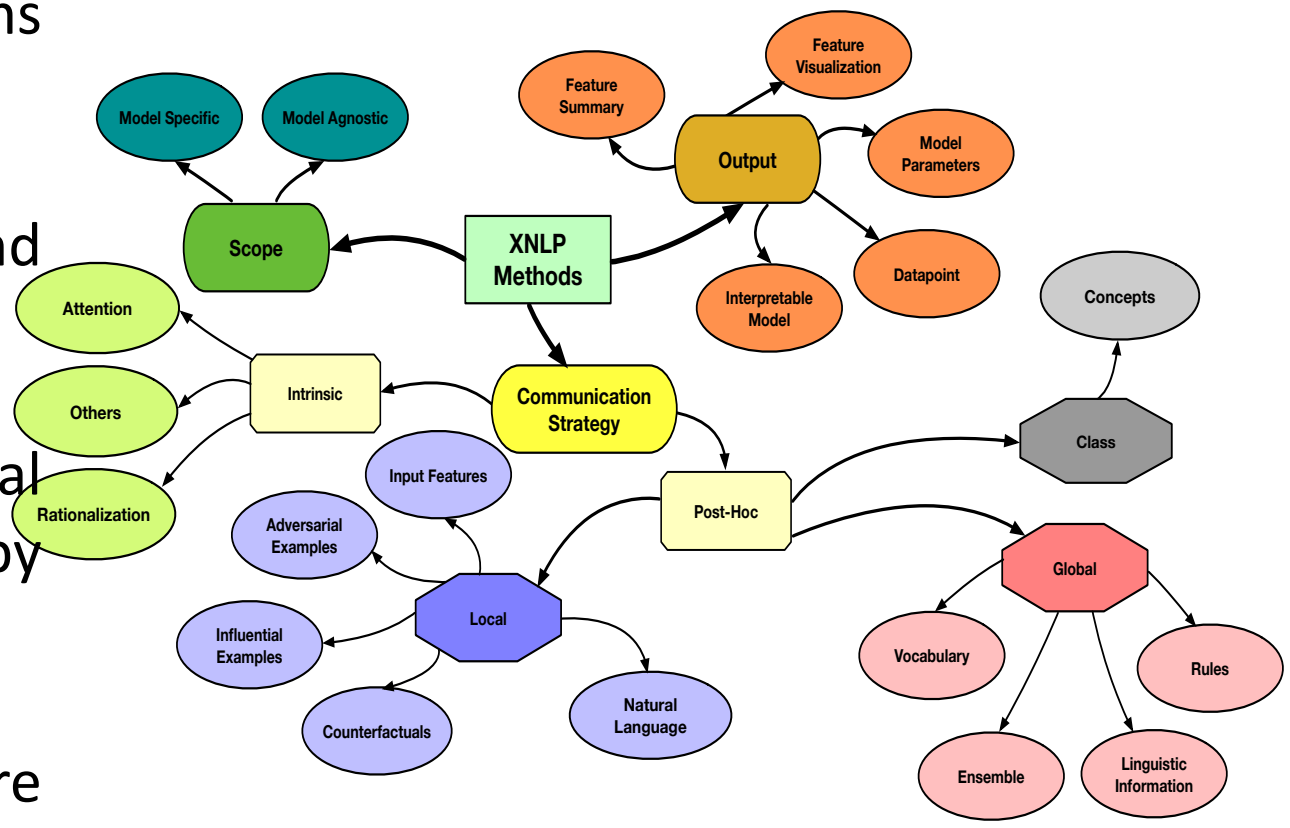
Model Name	District-wise		All Districts	
	Acc.	F1	Acc.	F1
Doc2Vec + SVM	0.72	0.69	0.79	0.77
Doc2Vec + XGBoost	0.68	0.59	0.67	0.57
IndicBert-First 512	0.65	0.62	0.73	0.71
IndicBert-Last 512	0.62	0.60	0.78	0.76
TF-IDF+IndicBert	0.76	0.74	**0.82**	**0.81**
TextRank+IndicBert	0.76	0.74	0.82	0.81
Salience Pred.+IndicBert	0.76	0.74	0.80	0.78
Multi-Task	**0.78**	**0.77**	*0.80*	*0.78*

Results

- There is a trade-off between explainability and prediction.
- Legal Domain requires explainable models to be usable by legal practitioners
- Read the paper for more details: https://aclanthology.org/2021.acl-long.313/
- Code Repository: https://github.com/Exploration-Lab/HLDC

XNLP Summary

- Explainability is a must for the systems deployed in the wild
- XNLP field has exploded in past few years and large number of methods are available.
- Rationalization methods provide natural language explanations and hence preferred by humans.
- We looked at Legal NLP domain where explainability is of paramount importance.



SemEval Task 6: LegalEval

https://sites.google.com/view/legaleval/