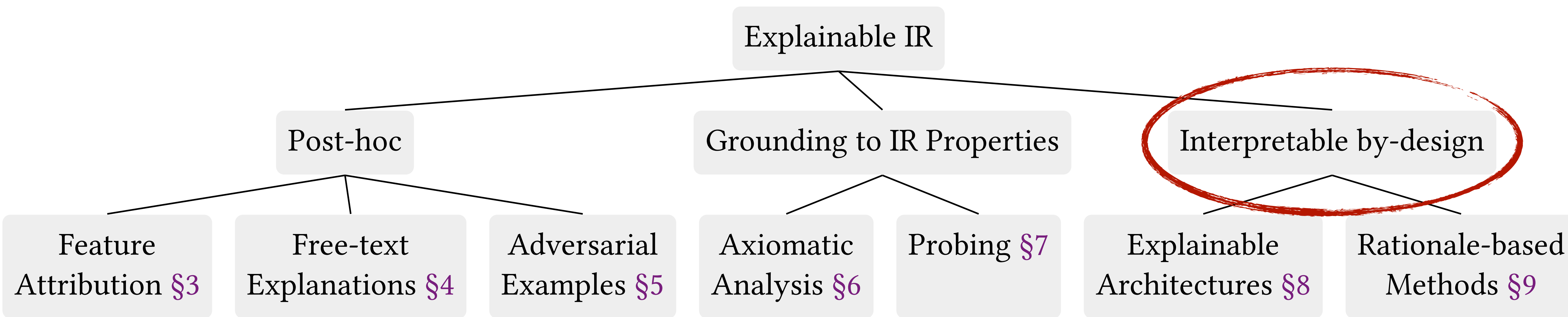
Explainable-by-design approaches

Explainable Information Retrieval

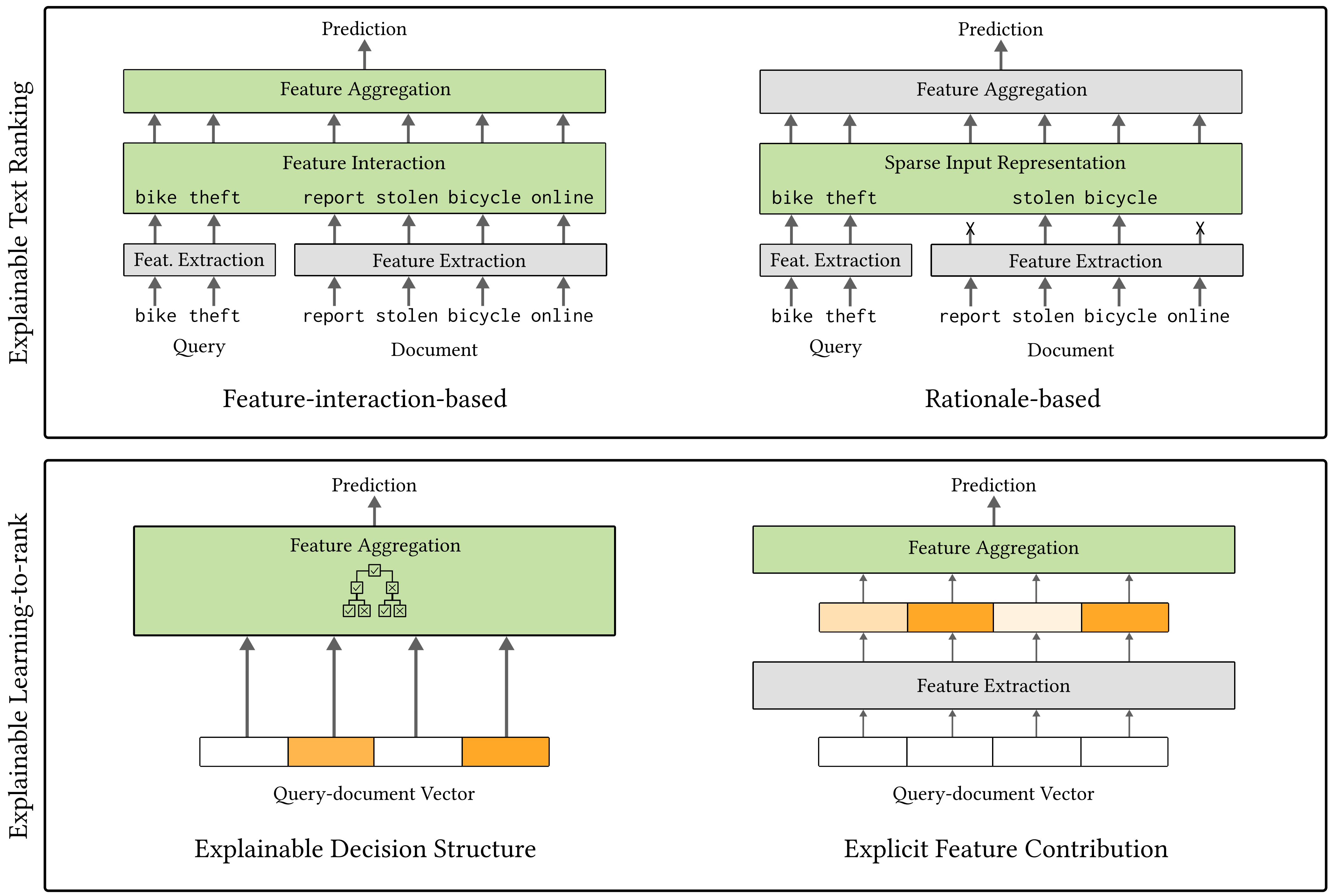
Interpretability Landscape

Explainable Information Retrieval: A Survey https://arxiv.org/abs/2211.02405

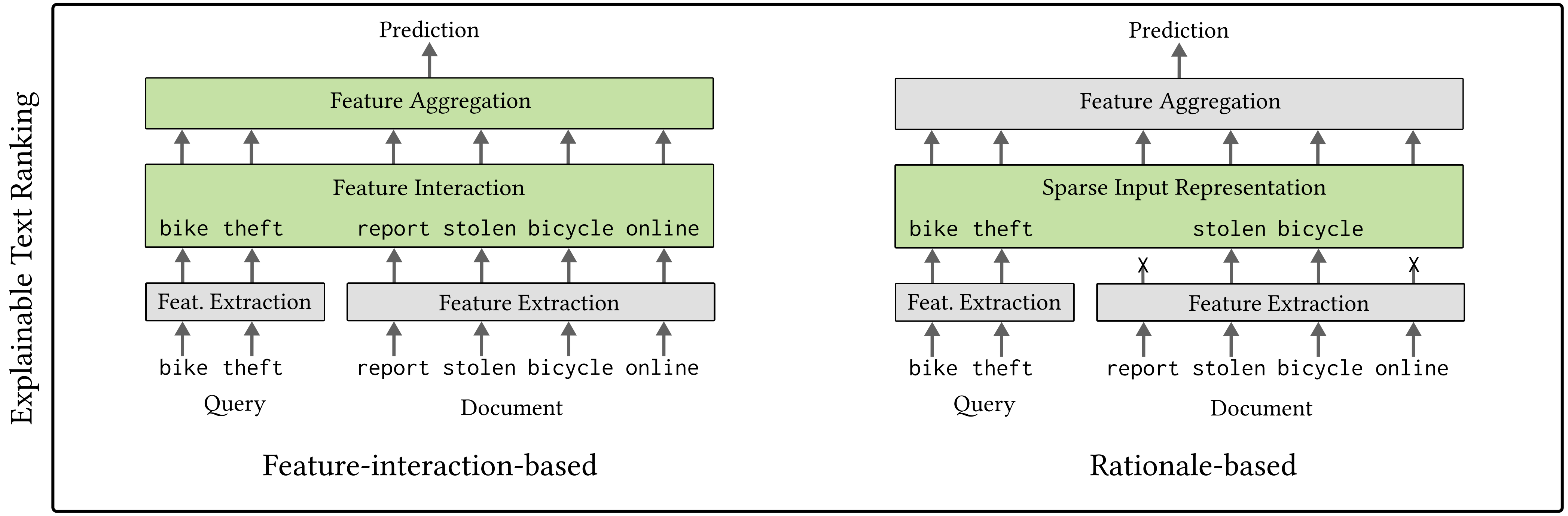
AVISHEK ANAND and LIJUN LYU, Delft University of Technology, The Netherlands
MAXIMILIAN IDAHL, YUMENG WANG, JONAS WALLAT, and ZIJIAN ZHANG, L3S Research Center, Leibniz University Hannover, Germany



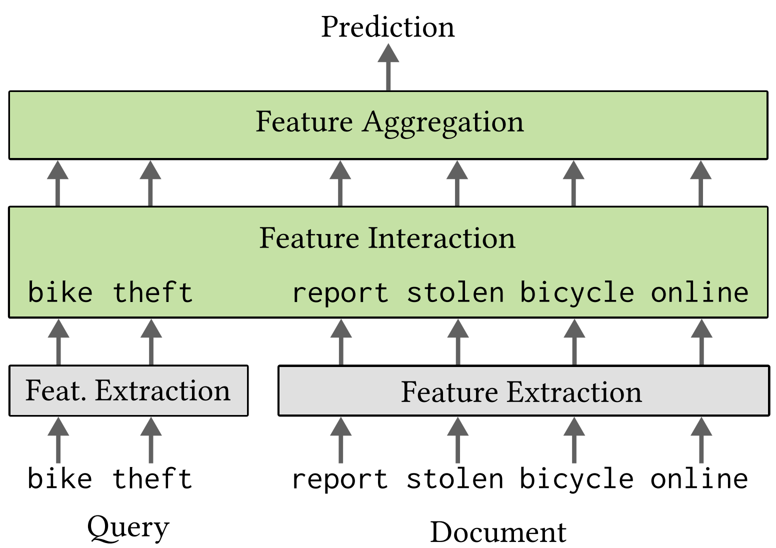
IBD Approaches



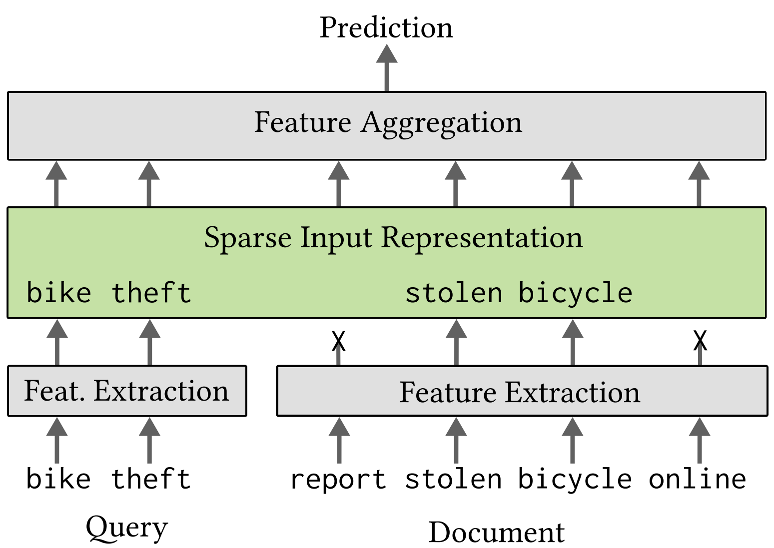
Text Based approaches



Text Based approaches



Feature-interaction-based



Rationale-based

What component is interpretable ?

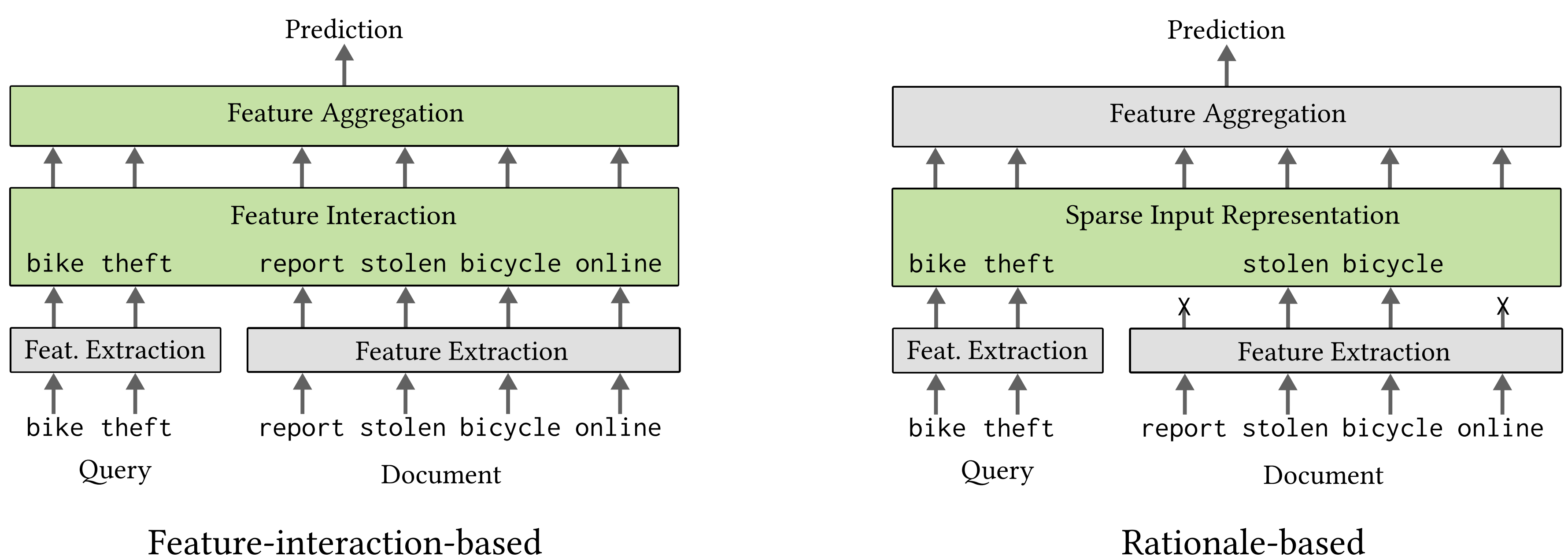
Feature extraction

Intermediate input representations

Feature Interaction and aggregation

Text Based approaches

BERT-based

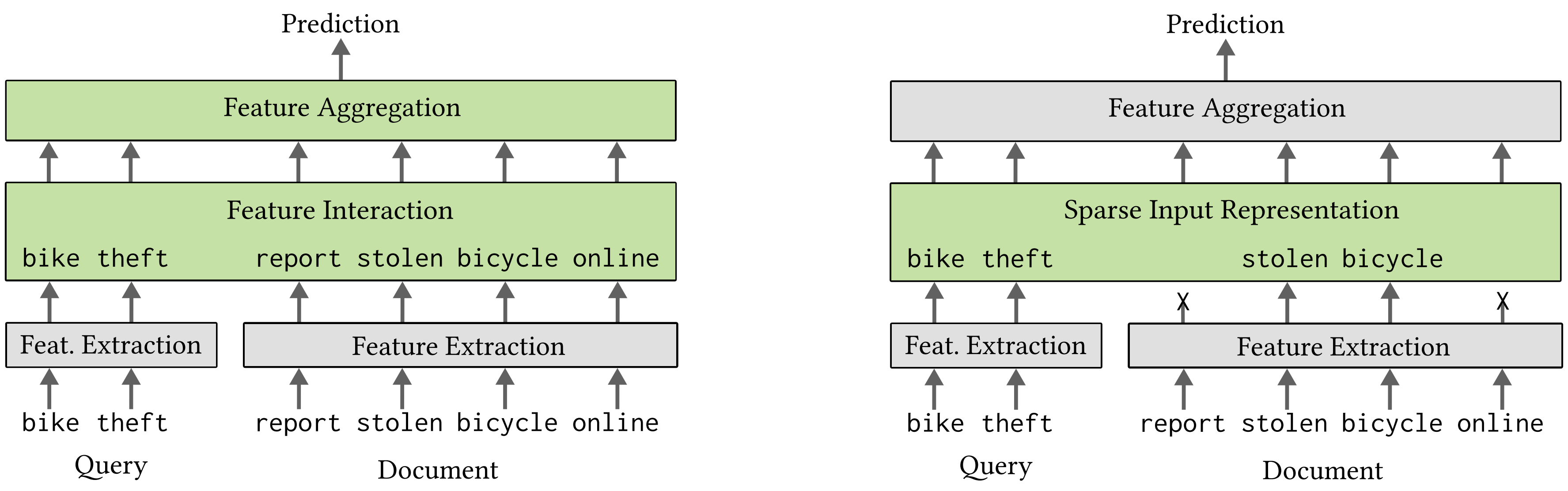


BERT-based

What component is interpretable ?

Feature extraction	Intermediate input representations	Feature Interaction and aggregation
Non-interpretable	interpretable	interpretable

Text Based approaches



Feature-interaction-based

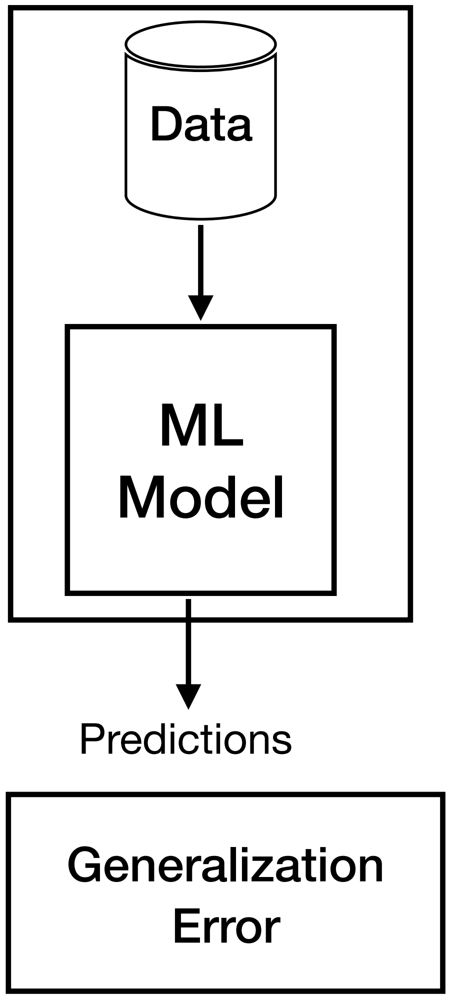
Rationale-based

What component is interpretable ?

Feature extraction	Intermediate input representations	Feature Interaction and aggregation
Non-interpretable	interpretable	interpretable

Standard Learning Setup

Standard ML



The movie experience was awful

Predict

Parameterised Model (BERT)

Explain then Predict

The movie experience was awful

Explain

Parameterised Model (BERT)

The movie experience **was awful**

Ensure prediction is solely on the explanations

Predict

Parameterised Model (BERT)

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

Optimizing explain then predict

The movie experience was awful

Explain

The movie experience **was awful**

Predict

$$\min_{\theta_e,\theta_g} \sum_{(\mathbf{x},\mathbf{y})\in D} \mathbb{E}_{\mathbf{z}\sim\mathbf{gen}(\mathbf{x})} \left[\mathrm{cost}(\mathbf{z},\mathbf{x},\mathbf{y})\right]$$

Task loss   Explanation loss

$$\mathrm{cost}(\mathbf{z},\mathbf{x},\mathbf{y}) = \mathcal{L}(\mathbf{z},\mathbf{x},\mathbf{y}) + \Omega(\mathbf{z}).$$

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

Sparsity   Continuity

Lei et. al [ACL 2018]

Optimizing explain then predict

The movie experience was awful

Explain

Parameterised Model (BERT)

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

The movie experience **was awful**

Predict

Parameterised Model (BERT)

$$\min_{\theta_e,\theta_g} \sum_{(\mathbf{x},\mathbf{y})\in D} \mathbb{E}_{\mathbf{z}\sim\mathbf{gen}(\mathbf{x})}\left[\mathrm{cost}(\mathbf{z},\mathbf{x},\mathbf{y})\right]$$

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

Doubly stochastic gradient / policy gradients/ REINFORCE

Explanation Performance

How human-like are the explanations ?

Explanation accuracy — Macro Token-wise F1

Fact Checking

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

Human annotation: 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. The combined statistical area of the region is the second largest in california after the Los Angeles area.

Extractive explanation: 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. The combined statistical area of the region is the second largest in california after the Los Angeles area.

Soft-matching metric: Token-wise precision, recall, and F1

Explanation Performance

How human-like are the explanations ?

Explanation accuracy — Macro Token-wise F1

Fact Checking

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

Human annotation: 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. The combined statistical area of the region is the second largest in california after the Los Angeles area.

Extractive explanation: 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. The combined statistical area of the region is the second largest in california after the Los Angeles area.

How much does Task Performance drop ?

Task accuracy — Macro F1

Benchmarks

ERASER: A Benchmark to Evaluate Rationalized NLP Models

Jay DeYoung[⋆Ψ], Sarthak Jain[⋆Ψ], Nazneen Fatema Rajani[⋆Φ], Eric Lehman[Ψ], Caiming Xiong[Φ], Richard Socher[Φ], and Byron C. Wallace[Ψ]

Name	Size (train/dev/test)	Tokens	Comp?
Evidence Inference	7958 / 972 / 959	4761	◇
BoolQ	6363 / 1491 / 2817	3583	◇
Movie Reviews	1600 / 200 / 200	774	◆
FEVER	97957 / 6122 / 6111	327	✓
MultiRC	24029 / 3214 / 4848	303	✓
CoS-E	8733 / 1092 / 1092	28	✓
e-SNLI	911938 / 16449 / 16429	16	✓

How human-like are the explanations ?

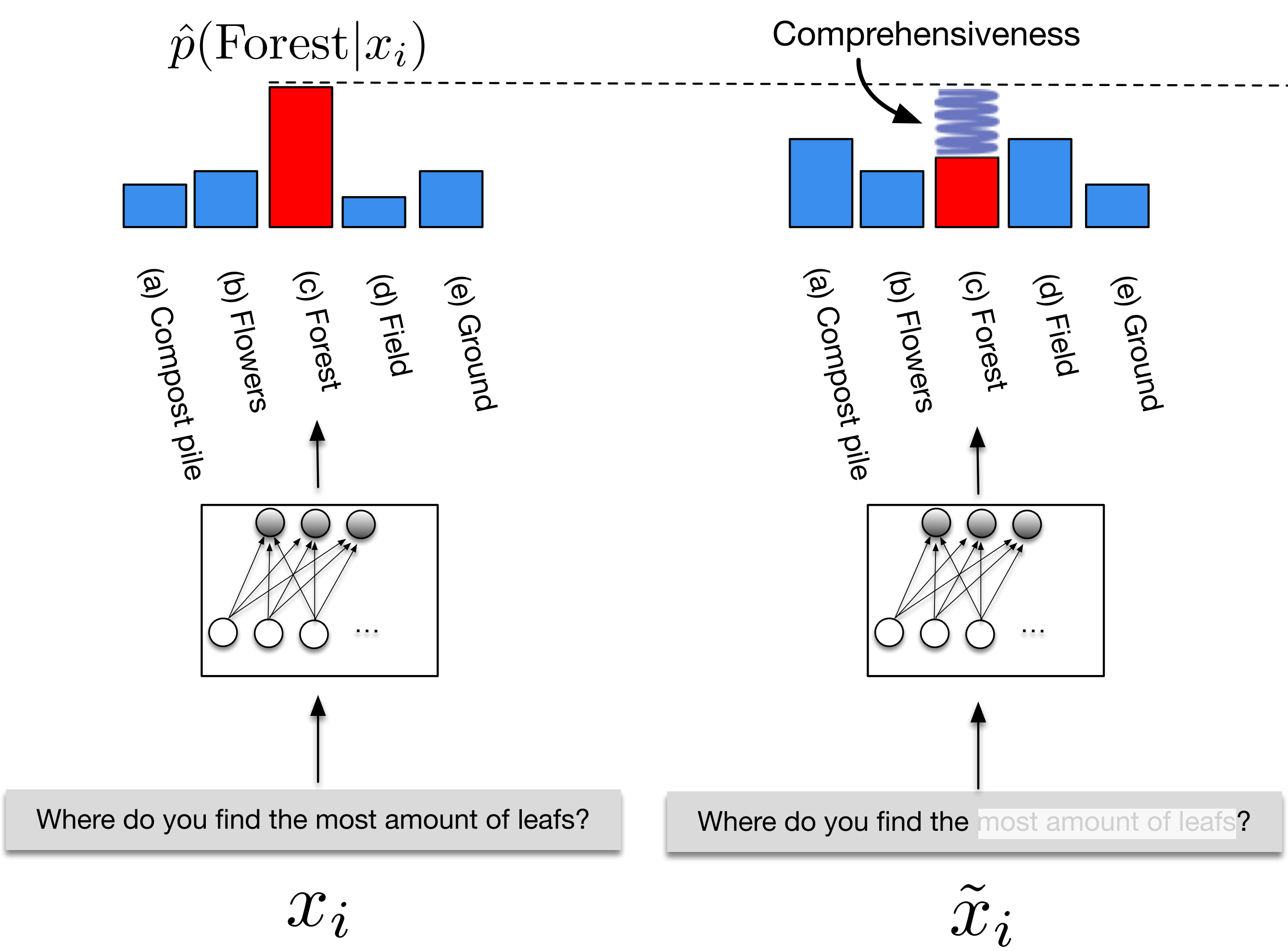
Soft-matching metric: Token-wise precision, recall, and F1

problem : a model may provide rationales that are plausible (agreeable to humans) but that it did not rely on the for its output.

Need: rationales extracted for an instance in this case ought to have meaningfully in- fluenced its prediction for the same

How faithful are the explanations to the model ?

Faithfulness



$$\text{comprehensiveness} = m(x_i)_j - m(x_i \backslash r_i)_j$$

Original pred. pred. with rationale removed

$$\text{sufficiency} = m(x_i)_j - m(r_i)_j$$

Original pred. pred. with just rationale

Problem

The movie experience was awful

Explain

Parameterised Model (BERT)

The movie experience was awful

Predict

Parameterised Model (BERT)

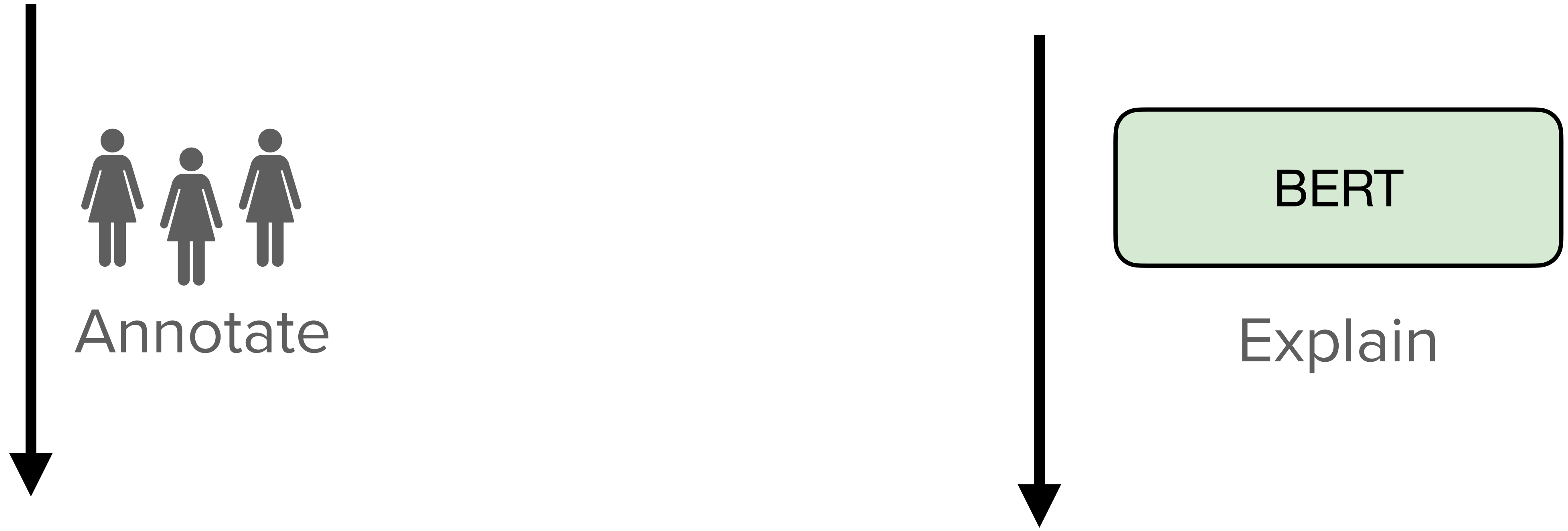
Optimizing just from the task labels is hard

Explanation generator is task unaware

Policy-gradient optimization known to be high variance

Explanation Data

The movie experience was awful



The movie experience was awful

0 0 0 1 1

Explanation Data

The movie experience was awful

Annotate

BERT

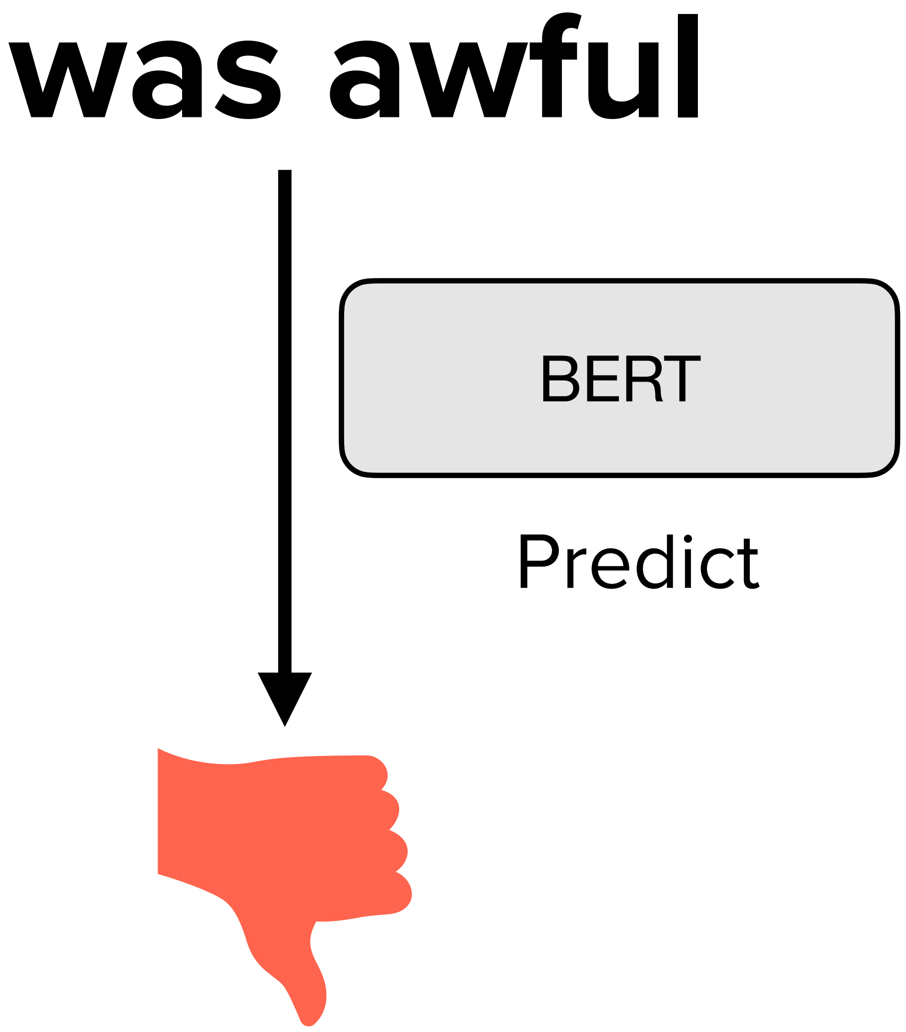
Explain

The movie experience **was awful**

0 0 0 1 1

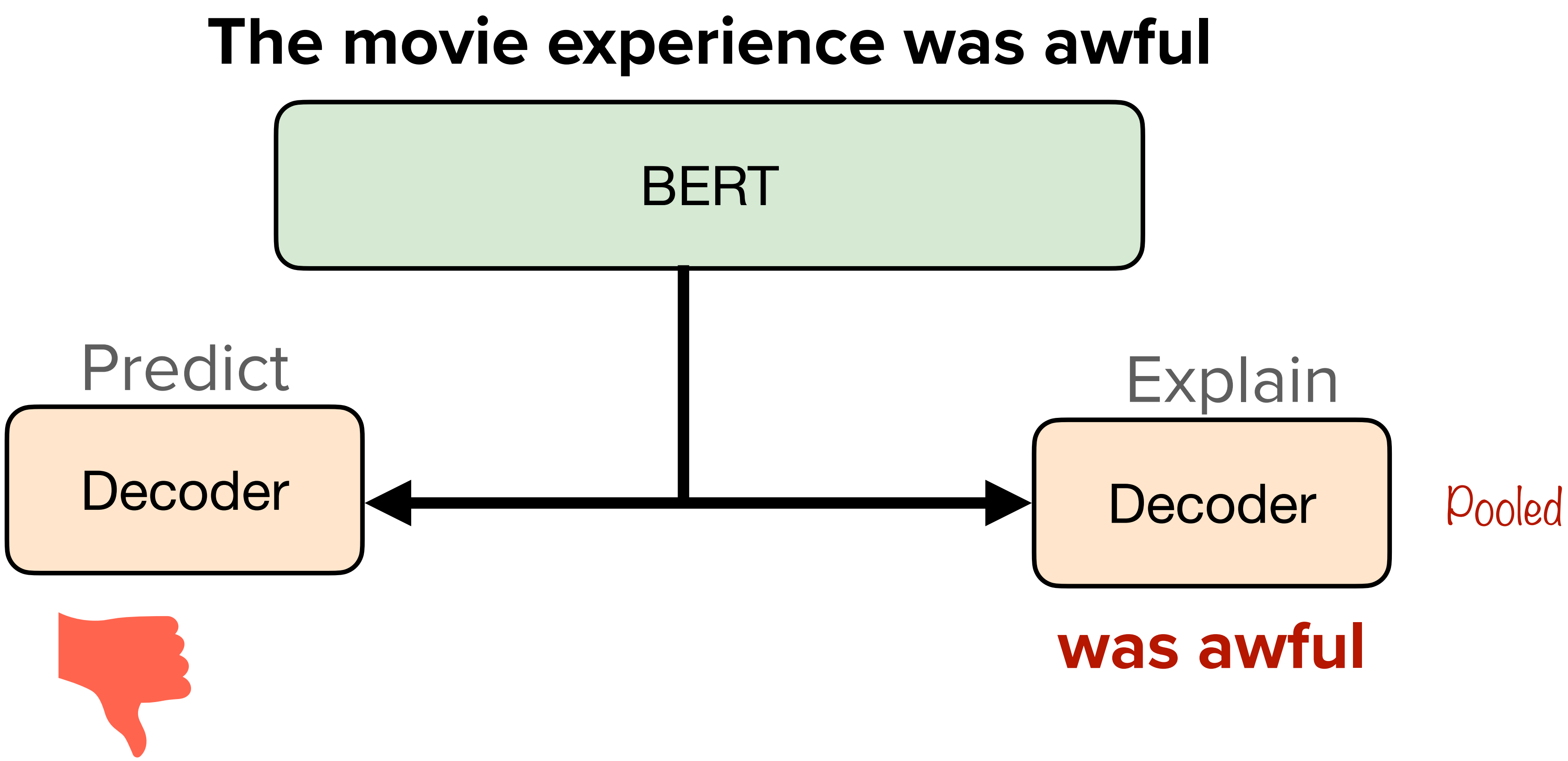
$$\mathcal{L}_{\text{exp}} = \frac{1}{|S|} \sum_{i=1}^{|S|} |S_{t^i}| \cdot \mathrm{BCE}\left(p^i, t^i\right)$$

Predict Model



Explain and Predict

Shared parameters during input encoding ensures that explanations are **task aware**



Multi-task Learning

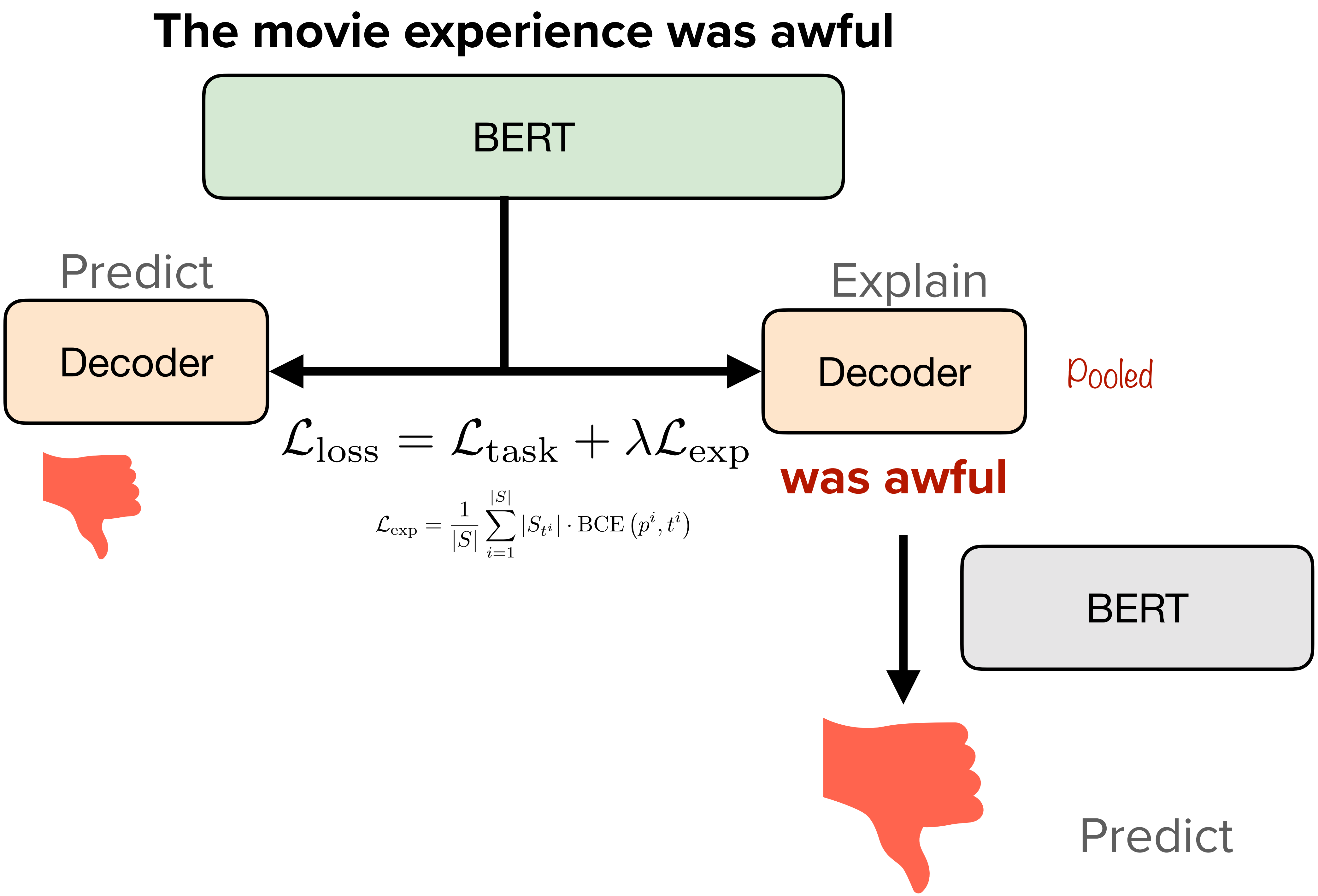
$$\mathcal{L}_{\text{loss}} = \mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{exp}}$$ Enforce sparcity

$$\mathcal{L}_{\text{exp}} = \frac{1}{|S|} \sum_{i=1}^{|S|} |S_{t^i}| \cdot \mathrm{BCE}\left(p^i, t^i\right)$$

Encoder representations regularised by explanation data

Explain and Predict, then Predict Again

Shared parameters during input encoding ensures that explanations are **task aware**



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

$$\mathcal{L}_{\text{exp}} = \frac{1}{|S|} \sum_{i=1}^{|S|} |S_{t^i}| \cdot \text{BCE}\left(p^i, t^i\right)$$

Explanation Performance

How human-like are the explanations ?

Explanation accuracy — Macro Token-wise F1

	Fact Checking	Question Answering	Sentiment Classification
No Explanation Data	0.83	0.43	0.32
ExPred	**0.84**	**0.64**	**0.35**

How much does Task Performance drop ?

Task accuracy — Macro F1

	Fact Checking	Question Answering	Sentiment Classification
Full Input	**0.91**	**0.70**	0.89
No Explanation Data	0.83	0.65	0.79
ExPred	0.89	0.69	**0.91**

No Explanation Data

Baselines that also produces binary masks
[Lei et al. 17], [Bastings et al. 19, Lehman et al. 19, DeYoung ‘20]

Fact Checking

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

Retrieved Document: 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. The combined statistical area of the region is the second largest in california after the Los Angeles area.

Fact Checking

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

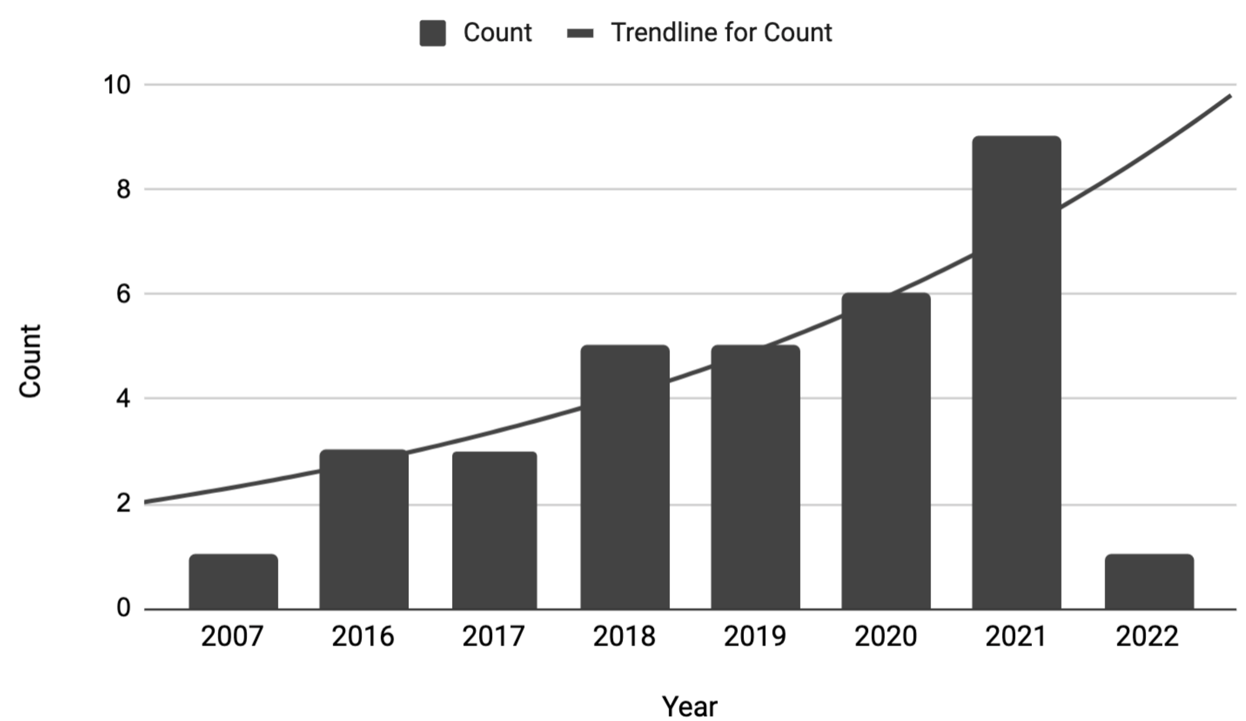
Retrieved Document: the san francisco bay area, referred to locally as the bay area is a populous region surrounding the san francisco and san pablo estuaries in northern california. The region encompasses the major cities and metropolitan areas of san jose, san francisco, and Oakland, along with smaller urban and rural areas. The bay area's nine counties are ......Santa Clara, Solana and Sonoma. The combined statistical area of the region is the second largest in california after the Los Angeles area.

[Zhang, Rudra & Anand WSDM '21]

Rationale-based approaches

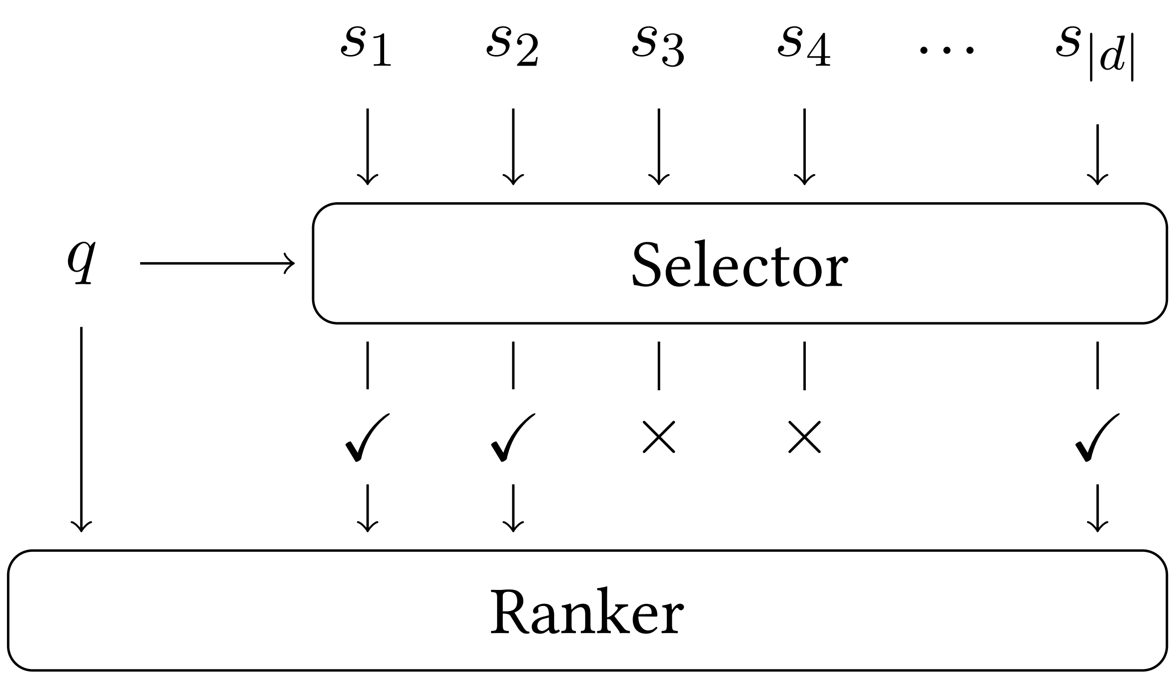
Rationalization for Explainable NLP: A Survey

SAI GURRAPU, Department of Computer Science, Virginia Tech, USA
AJAY KULKARNI, Department of Computational and Data Sciences, George Mason University, USA
LIFU HUANG, Department of Computer Science, Virginia Tech, USA
ISMINI LOURENTZOU, Department of Computer Science, Virginia Tech, USA
LAURA FREEMAN, Department of Statistics, Virginia Tech, USA
FERAS A. BATARSEH, Department of Electrical and Computer Engineering, Virginia Tech, USA



popular in NLP research

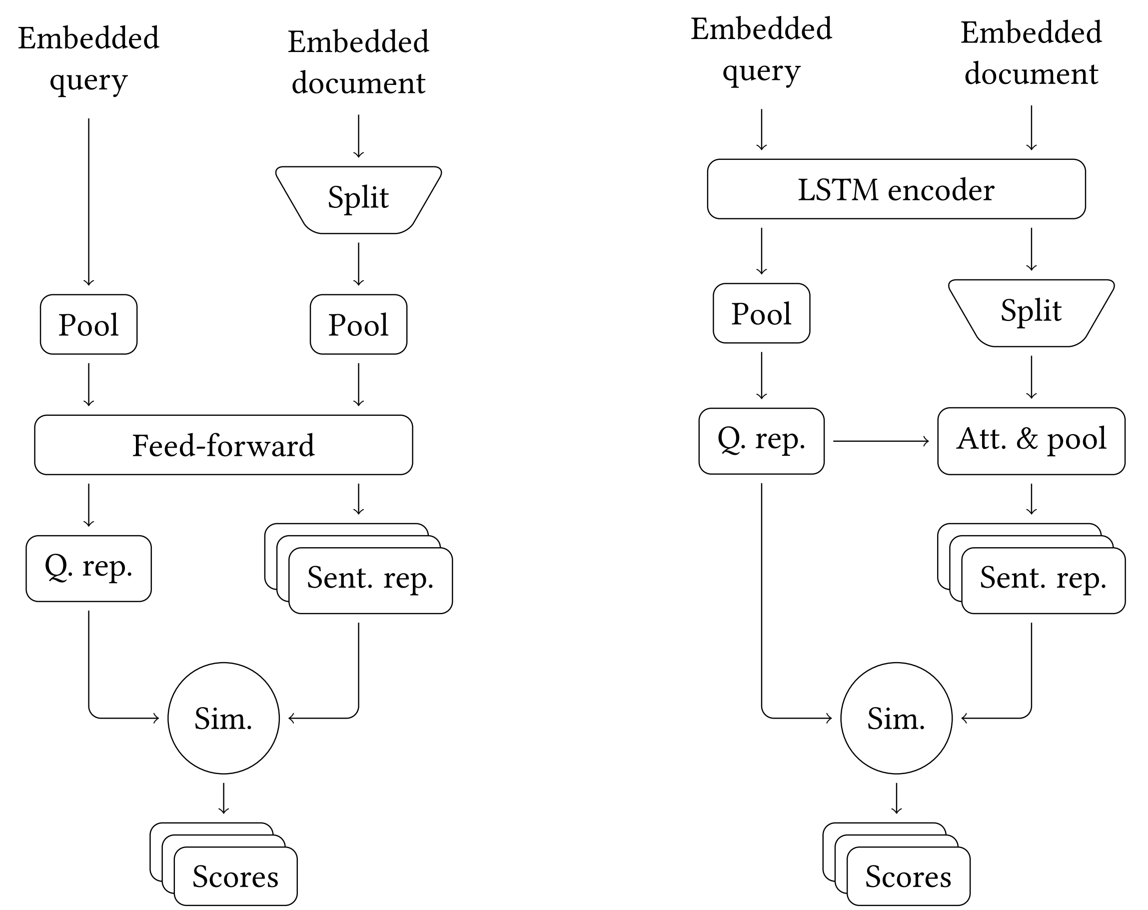
Rationales for ranking



Select and rank paradigm: Can we trade-off sparsity and ranking quality by controllably selecting a subset of sentences.

Extractive explanations for text ranking [Leonhardt, Rudra & Anand TOIS '23]

Selectors



(a) Linear selector **(b)** Attention-based LSTM selector

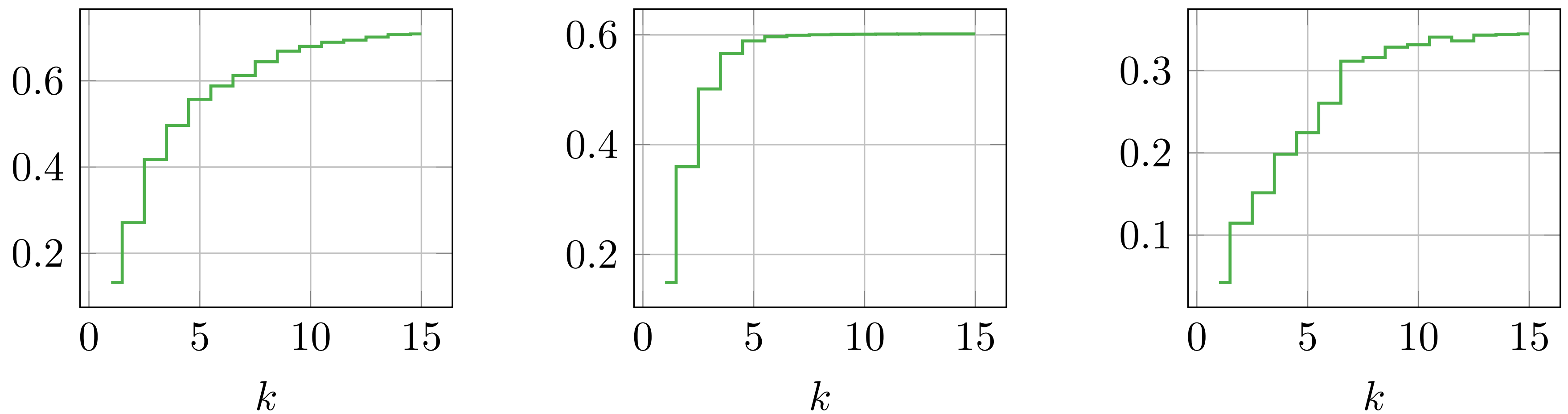
Selectors: Selectors should be simple for efficiency

Optimizing selectors: Gumbel-max trick + relaxed subset sampling

Extractive explanations for text ranking [Leonhardt, Rudra & Anand TOIS '23]

Insights

	TREC-DL-Doc'19			Core17			ClueWeb09		
	AP	nDCG@20	RR	AP	nDCG@20	RR	AP	nDCG@20	RR
QL	0.237	0.487[ab]	0.785	0.203	0.395	0.686	0.165	0.277	0.487
Doc-Labeled	0.203	0.434[ab]	0.731	0.237	0.437	0.742	0.165	0.284	0.503
BERT-3S	0.245	0.519[ab]	0.799	0.204	0.406	0.694	0.178	0.306	0.544
BERT-CLS	0.260	0.581	0.874	0.196	0.419	0.749	0.178	0.313	0.572
PL-SEM	0.265	0.571	0.920	0.207	0.414	0.768	0.167	0.286	0.534
[a] S&R-LIN	0.269	0.597	0.946	0.203	0.411	0.710	0.174	0.303	0.535
[b] S&R-ATT	0.271	0.590	0.924	0.205	0.403	0.714	0.168	0.292	0.518



(a) Fever **(b)** HotpotQA **(c)** SciFact

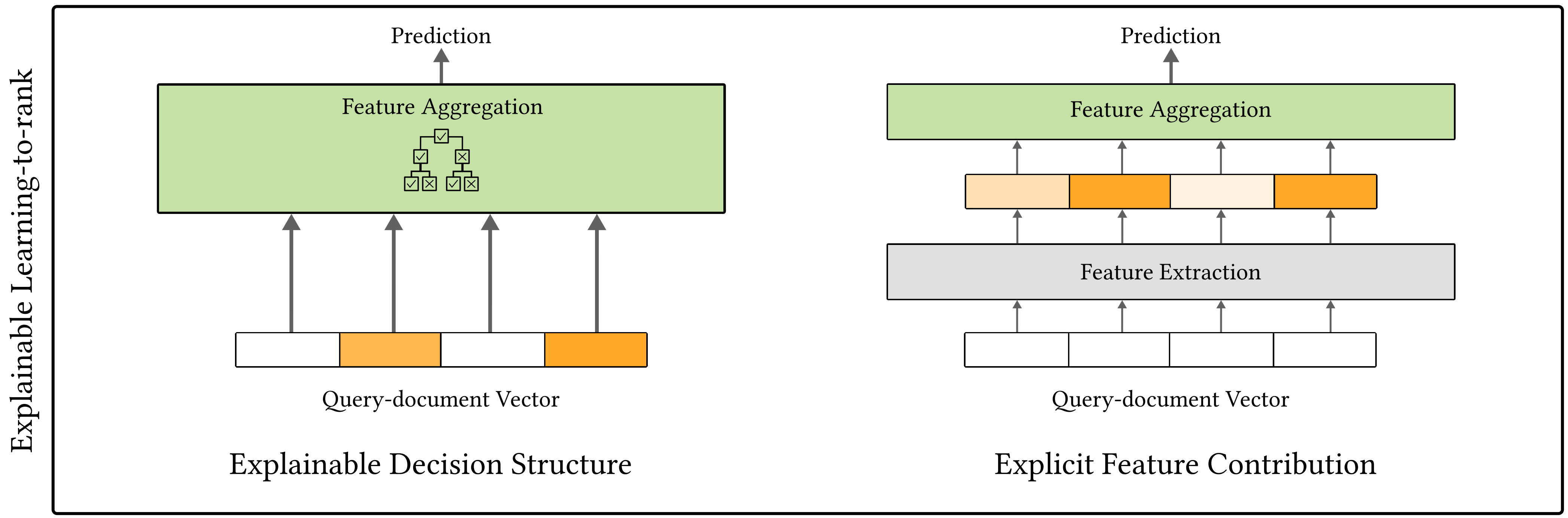
Extractive explanations for text ranking [Leonhardt, Rudra & Anand TOIS '23]

Extractive Explanations for Rankings

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

Retrieved Document: 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. The combined statistical area of the region is the second largest in california after the Los Angeles area.

Learning-to-rank approaches



GAMs

Learning to rank with Generalized additive models

Interpretable Ranking with Generalized Additive Models

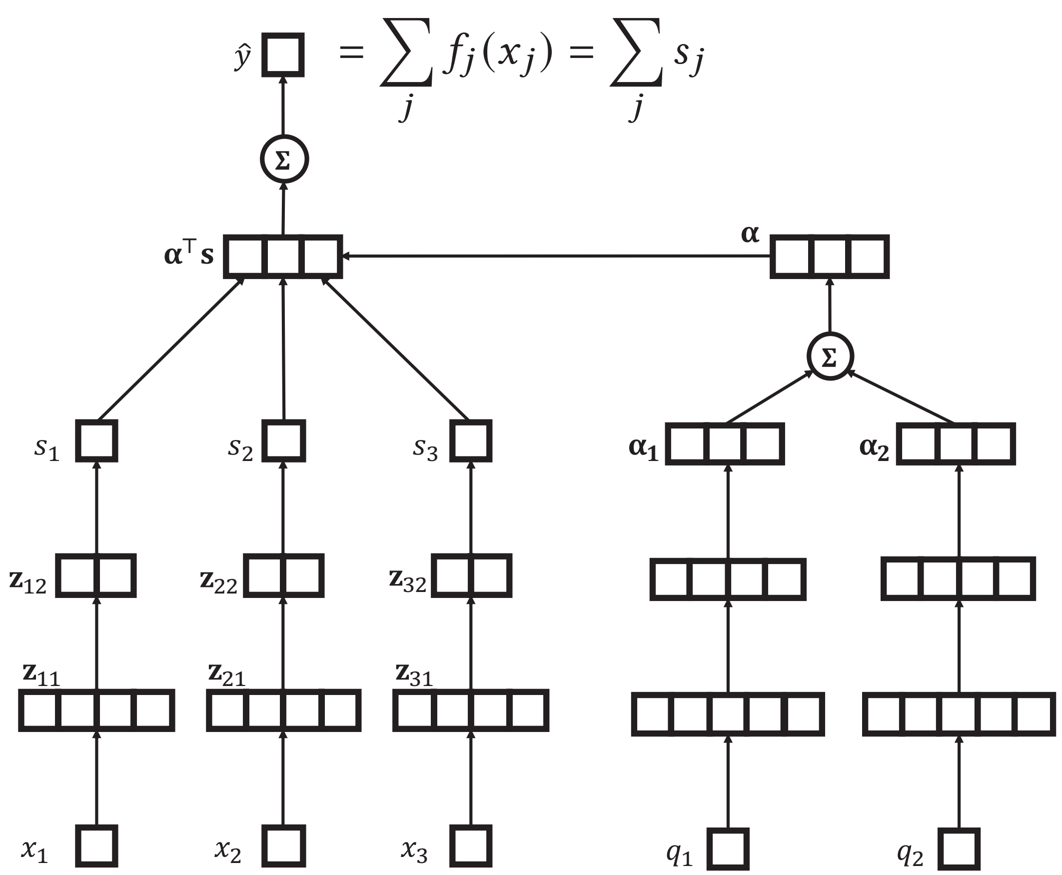
Honglei Zhuang, Xuanhui Wang, Michael Bendersky, Alexander Grushetsky, Yonghui Wu,
Petr Mitrichev, Ethan Sterling, Nathan Bell, Walker Ravina, Hai Qian
{hlz,xuanhui,bemike,grushetsky,yonghui,petya,esterling,nathanbell,walkerravina,hqian}@google.com
Google

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N} \qquad \mathbf{x}_i = (x_{i1}, \cdots, x_{in})$$

$$g(\hat{y}_i) = f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_n(x_{in})$$

A function for each feature

Ranking GAMs



$$f_j(x_j) = s_j = \mathbf{W}_j \mathbf{z}_{jH} + b_j$$

A neural net for each feature

$$\mathbf{z}_{j1} = \sigma(\mathbf{W}_{j1} x_j + \mathbf{b}_{j1})$$
$$\mathbf{z}_{j2} = \sigma(\mathbf{W}_{j2} \mathbf{z}_{j1} + \mathbf{b}_{j2})$$
$$\ldots$$
$$\mathbf{z}_{jH} = \sigma(\mathbf{W}_{jH} \mathbf{z}_{j(H-1)} + \mathbf{b}_{jH})$$

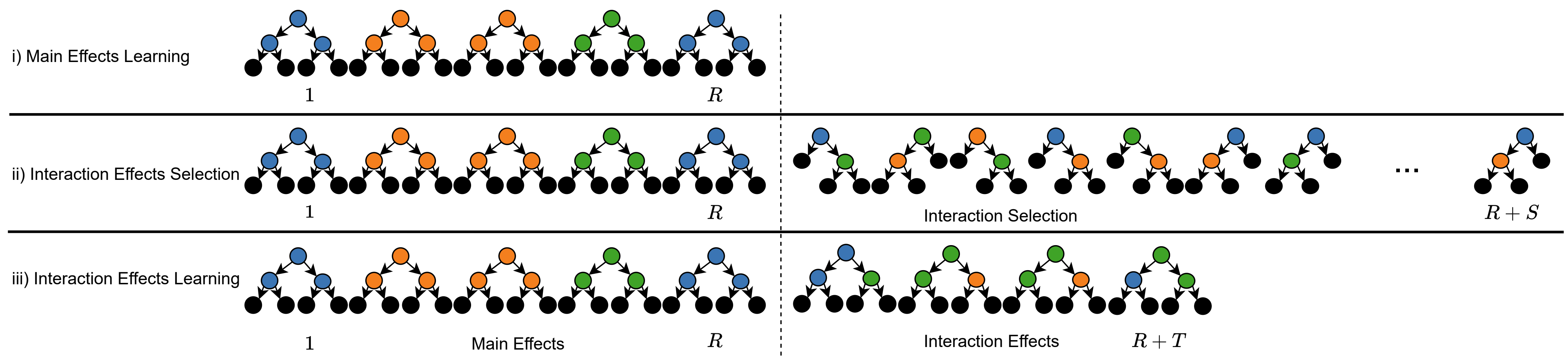
$$S^* = \underset{S=\{(x_k,y_k)\}_{k=1}^K}{\arg\min} \frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}} \left\| f(x_i) - PWL_S(x_i) \right\|^2$$

$$PWL_S(x) = \begin{cases} y_1 & \text{if } x < x_1, \\ \frac{y_{k+1}-y_k}{x_{k+1}-x_k}(x - x_k) + y_k & \text{if } x_k \le x \le x_{k+1}, \\ y_K & \text{if } x > x_K. \end{cases}$$

ILMART

Problem in GAMs : No interaction between features

$$\hat{y} = \overbrace{\underbrace{\sum_{j \in \mathcal{J}} \tau_j(x_j)}_{R \text{ trees}}}^{|\mathcal{J}|=p \text{ main effects}} + \overbrace{\underbrace{\sum_{(i,j) \in \mathcal{K}} \tau_{ij}(x_i, x_j)}_{T \text{ trees}}}^{|\mathcal{K}|=K \text{ interaction effects}}$$



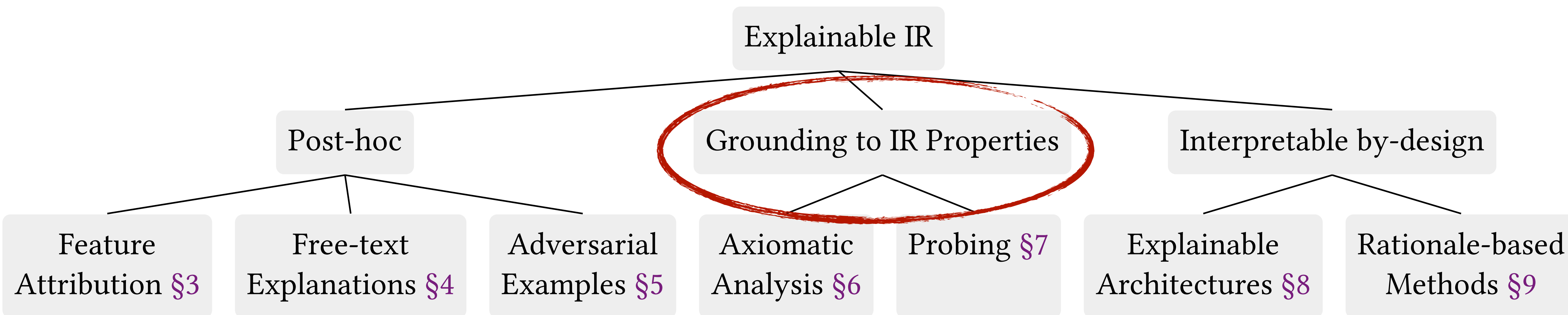
ILMART: Interpretable Ranking with Constrained LambdaMART [Lucchese et al. SIGIR '22]

Interpretability Landscape

Explainable Information Retrieval: A Survey

https://arxiv.org/abs/2211.02405

AVISHEK ANAND and LIJUN LYU, Delft University of Technology, The Netherlands
MAXIMILIAN IDAHL, YUMENG WANG, JONAS WALLAT, and ZIJIAN ZHANG, L3S Research Center, Leibniz University Hannover, Germany

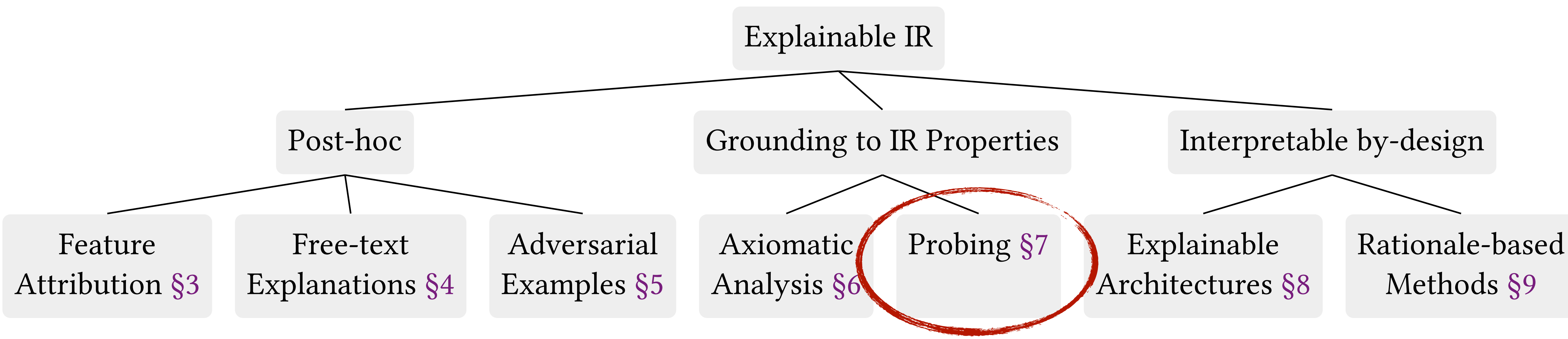


Interpretability Landscape

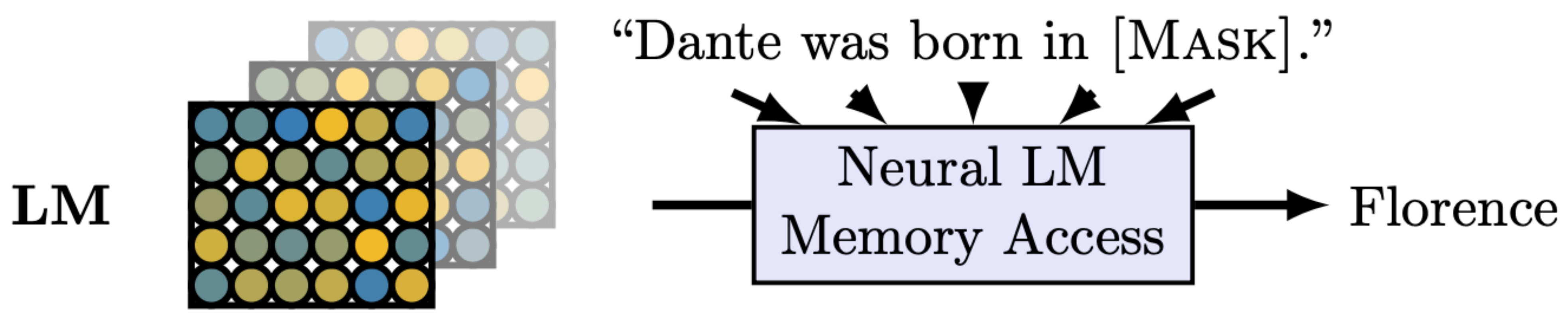
Explainable Information Retrieval: A Survey

https://arxiv.org/abs/2211.02405

AVISHEK ANAND and LIJUN LYU, Delft University of Technology, The Netherlands
MAXIMILIAN IDAHL, YUMENG WANG, JONAS WALLAT, and ZIJIAN ZHANG, L3S Research Center, Leibniz University Hannover, Germany



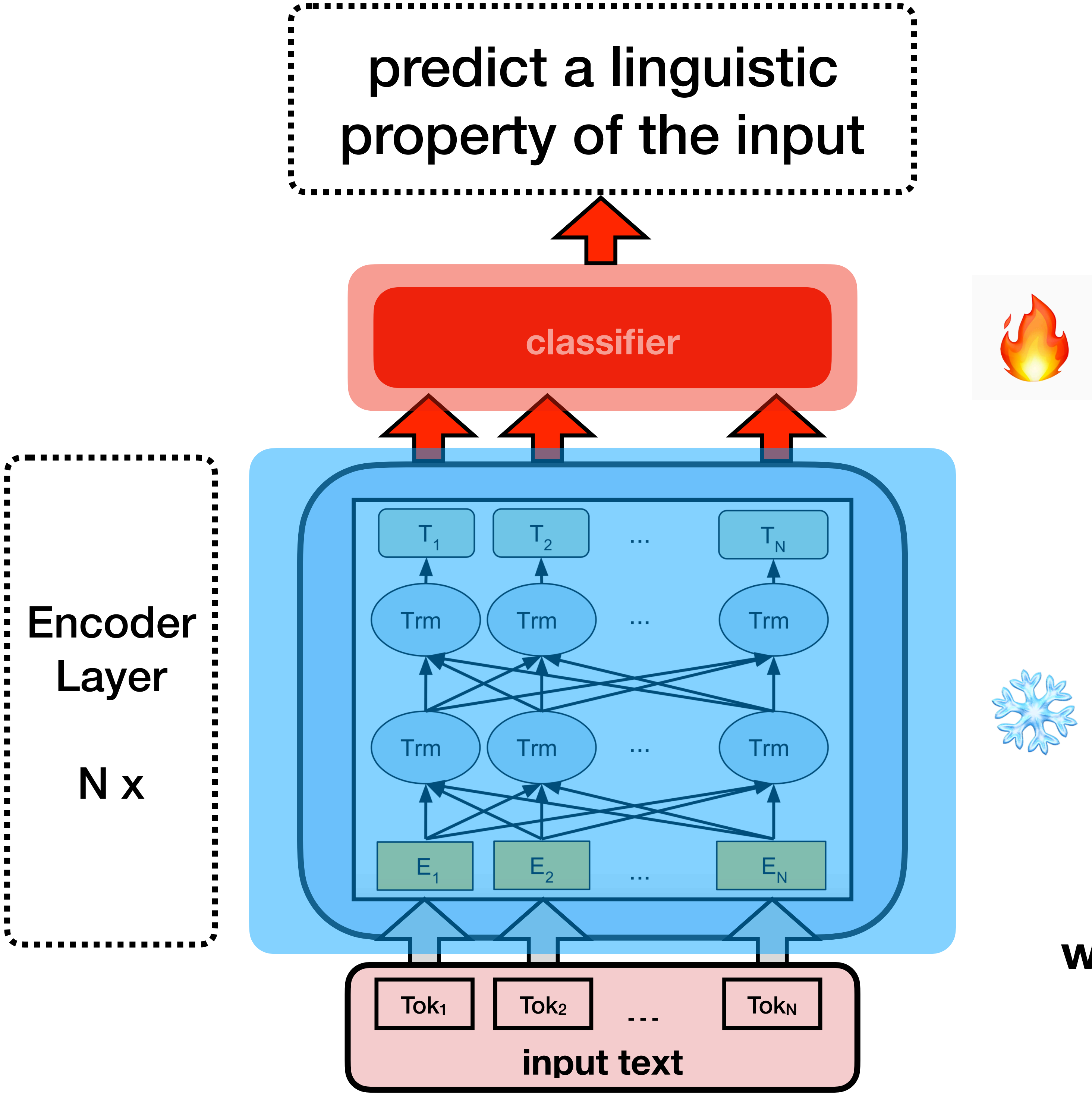
Probing Philosophy



If we can train a classifier to predict a property of the input text based on its representation, it means the property is encoded somewhere in the representation

[petroni et al. 2019]

How do we probe ?



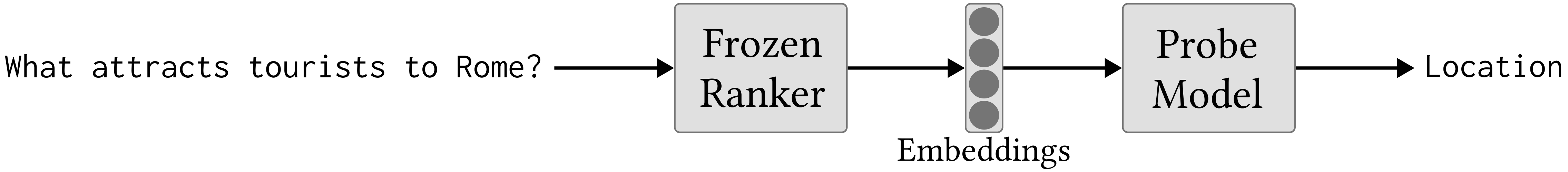
the classifier's weights are updated

train the classifier only

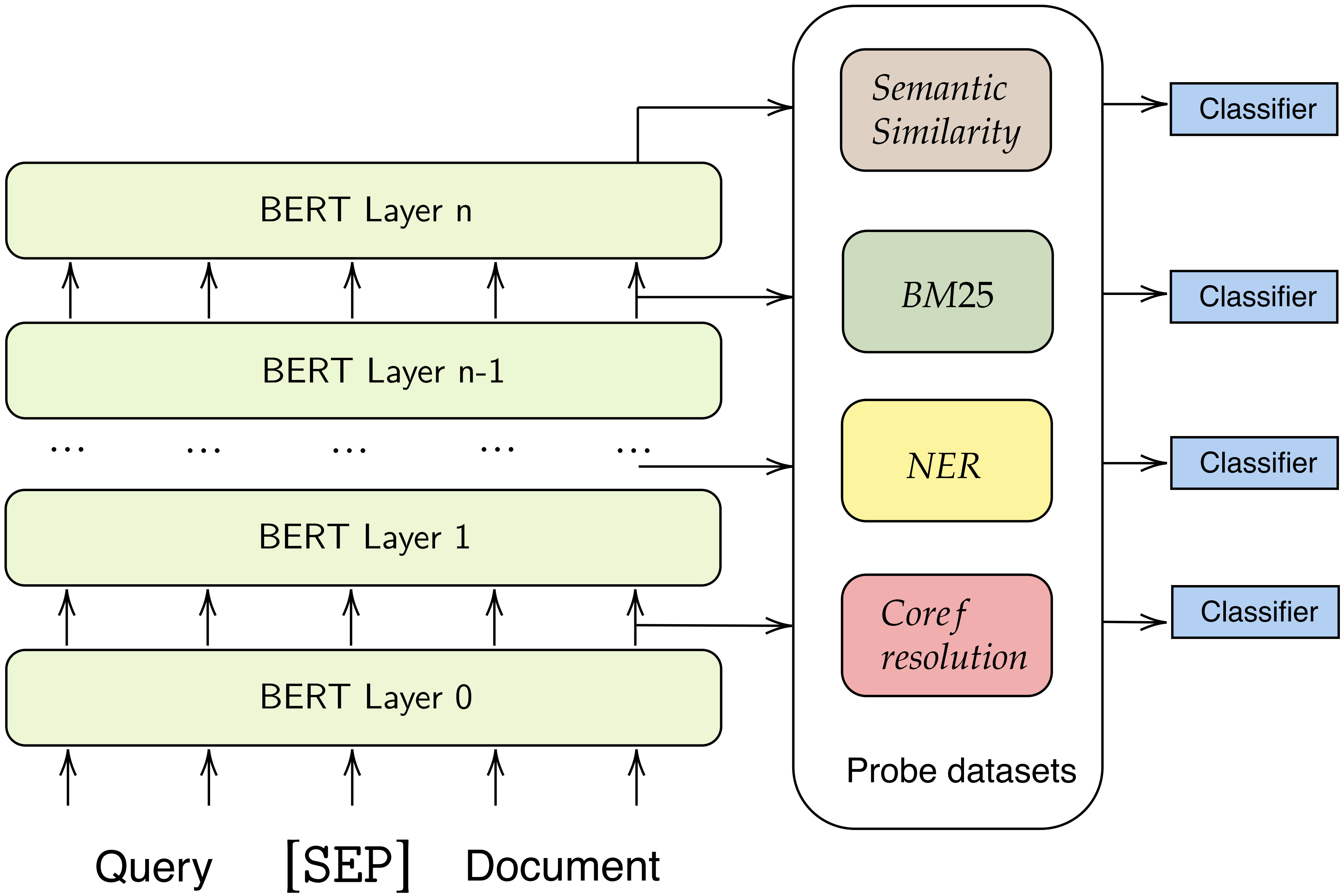
no further fine-tuning

the encoder's weights are fixed

Probing rankers



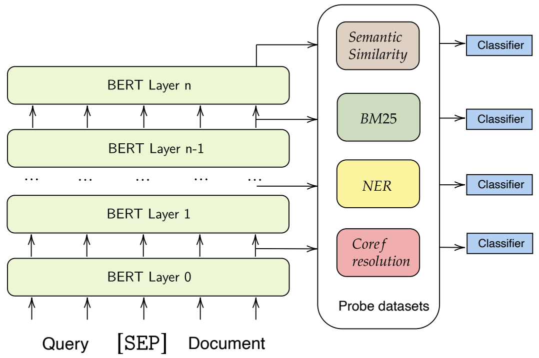
Probing Rankers

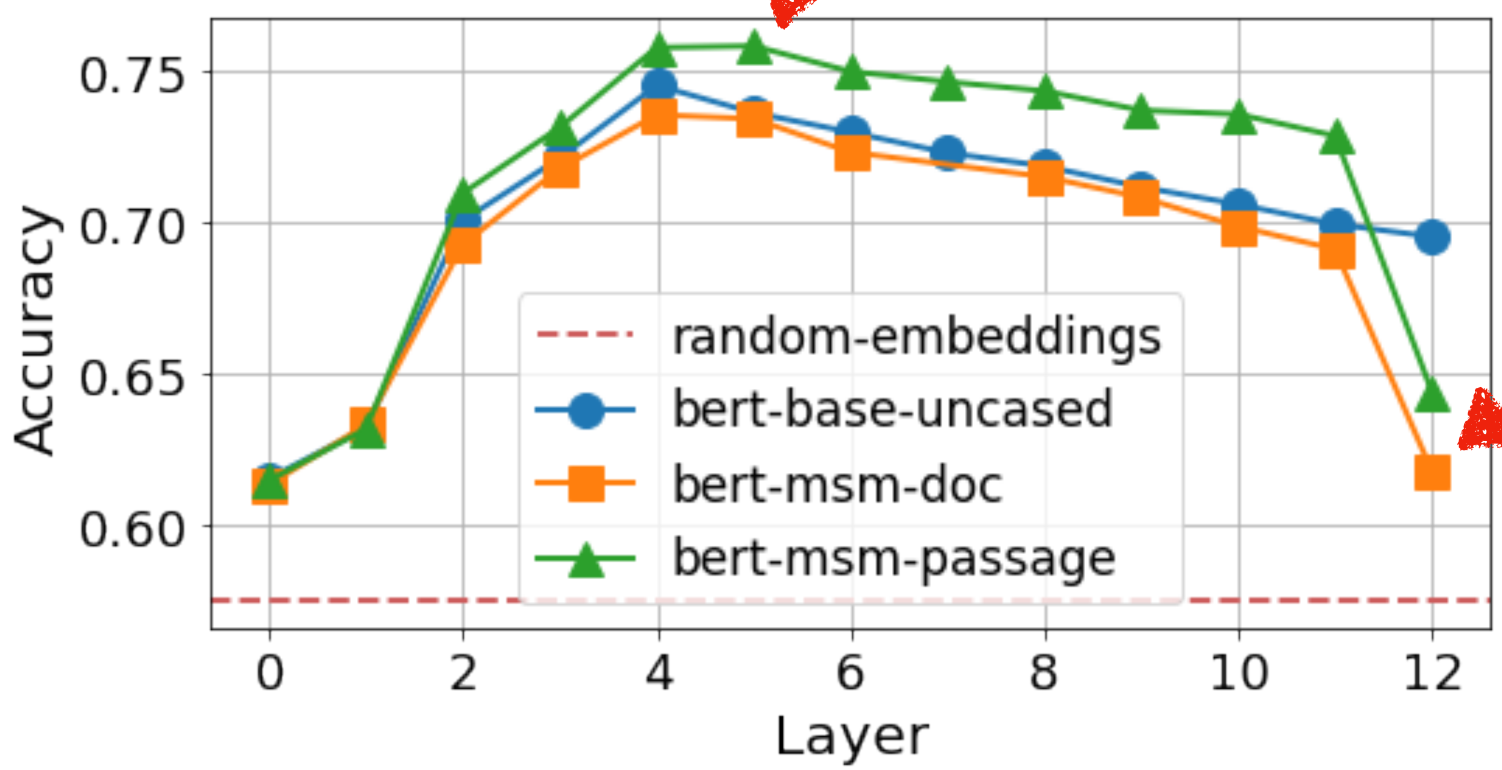


If the document representation can do well on a IR ability then it understands or exhibits that ability well…

[Wallat et al ECIR ’23]

Probing Rankers

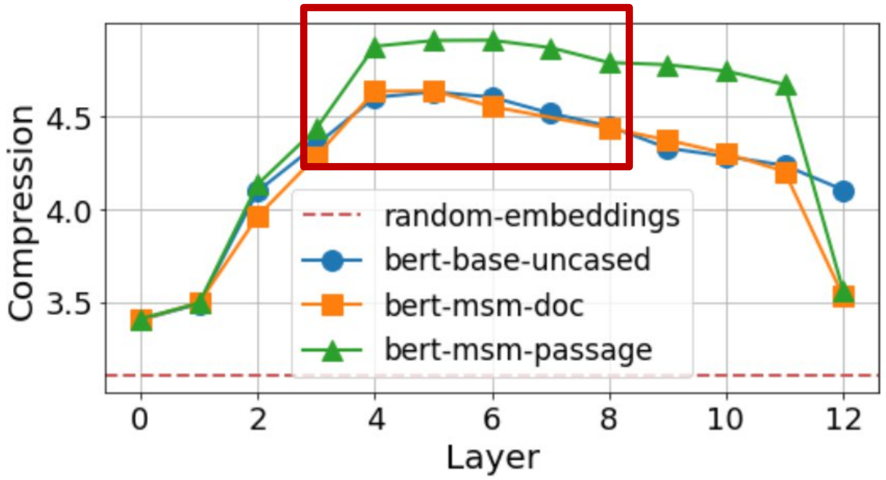




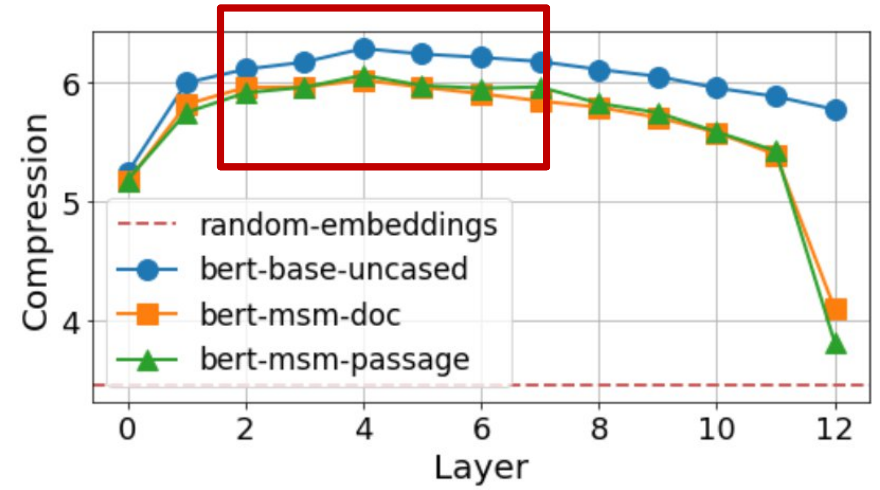
If the document representation can do well on a IR ability then it understands or exhibits that ability well…

[Wallat et al ECIR '23]

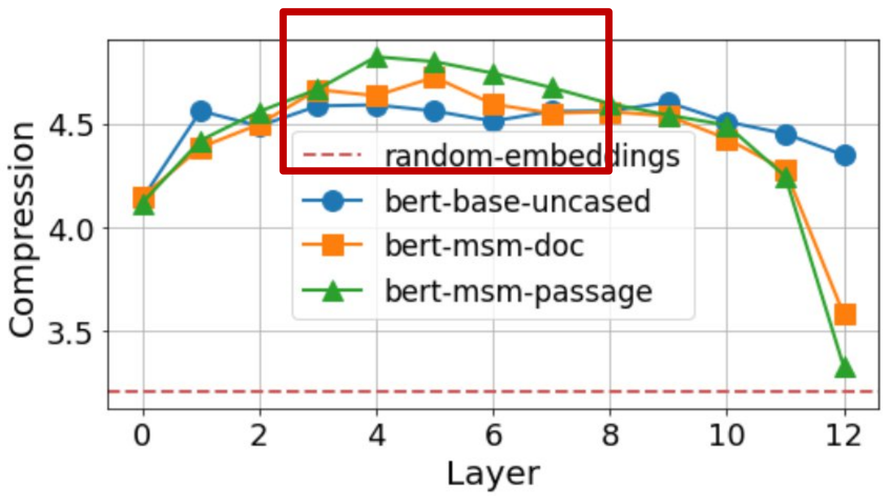
IR abilities in representation



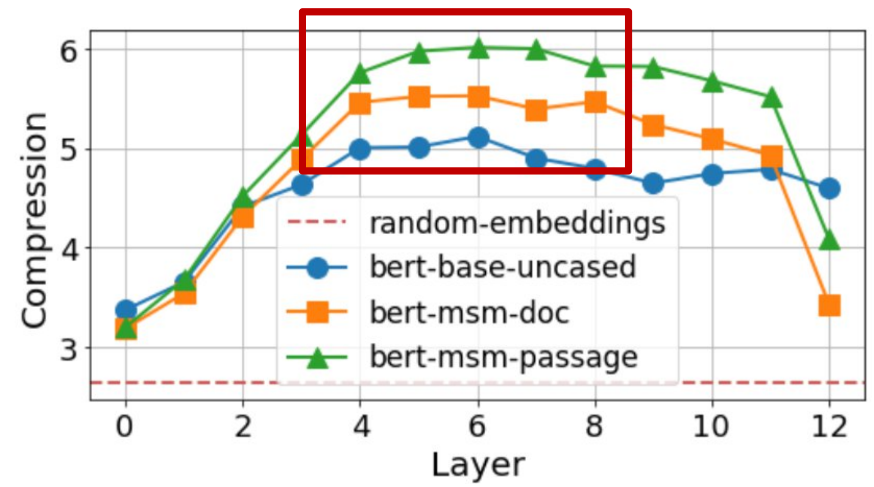
BM25



NER



Semantic Similarity



Coreference resolution