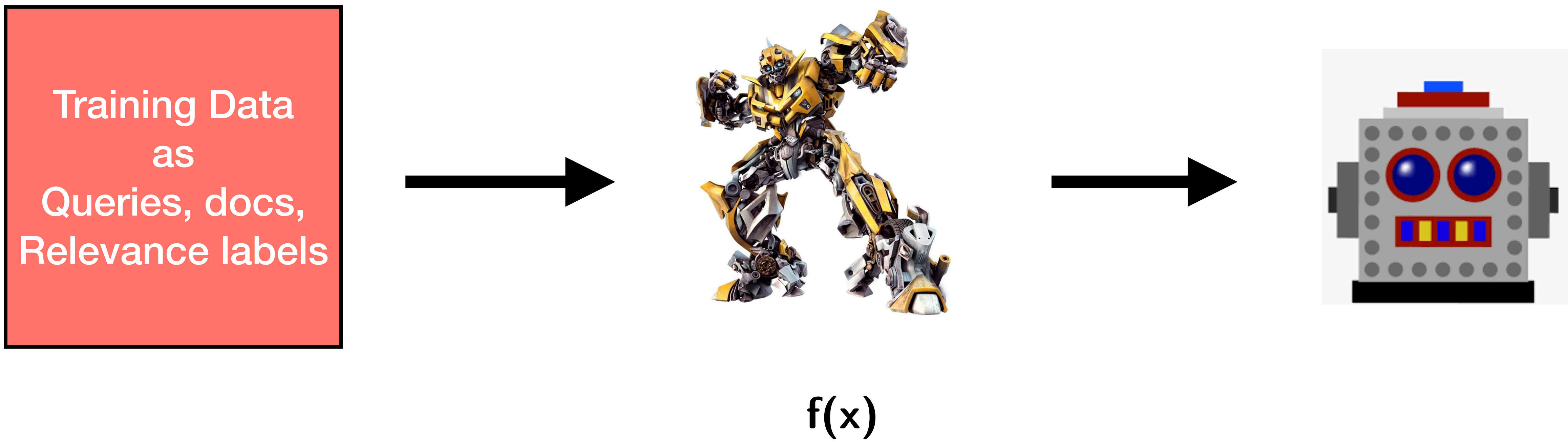
Posthoc Interpretability

Explainable Information Retrieval

Procheta Sen

University of Liverpool

Setting: Posthoc Interpretability



Approximate the learned model with a simpler understandable model

Feature Attribution Based Explanations

Free-Text Explanations

What is a good explanation ?

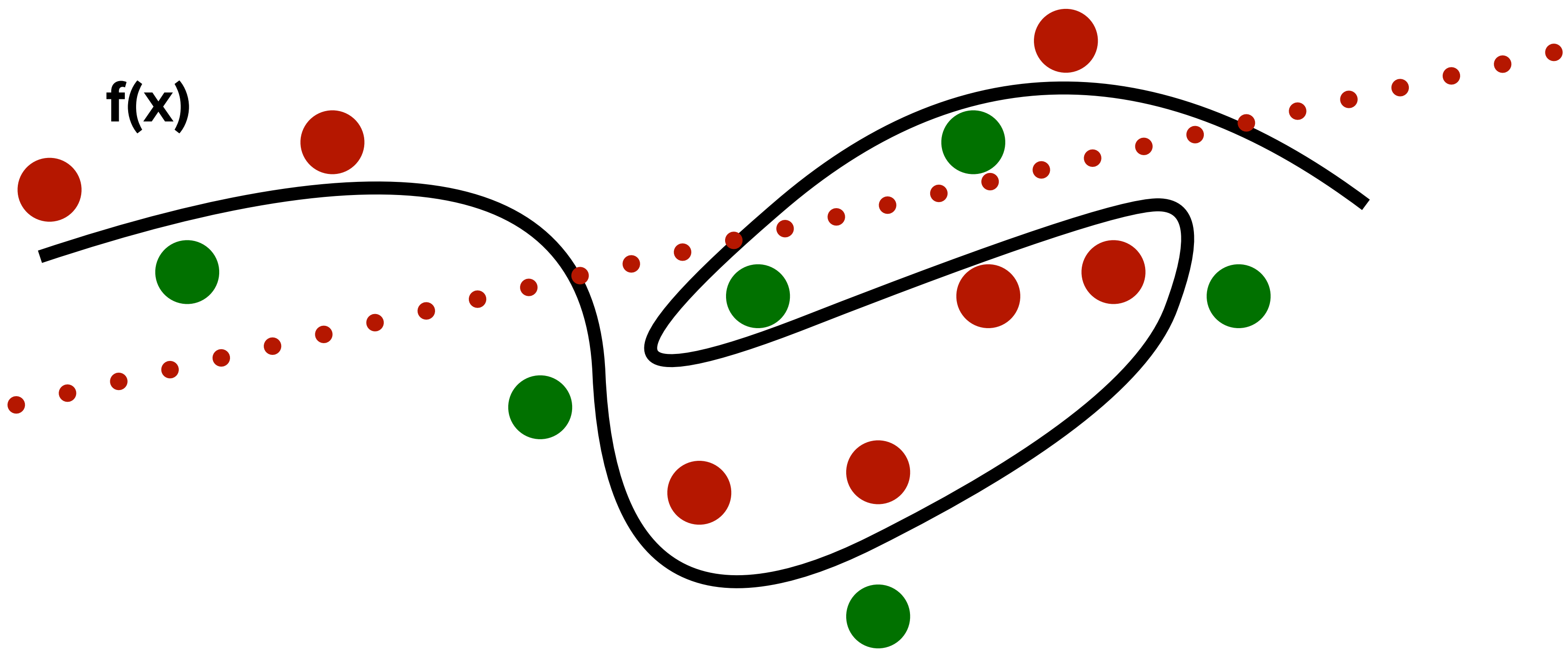
- **Accurate** — Should find the right reasons behind a decision
- **Fidelity** — Closely mimic the behaviour of the learnt model
- Explanation should be **understandable**
  - Explanation space — words, phrases,…
- Explanation model should also be **simple**
  - Linear model, BM25, ..

Categorisation of Explainable Approaches

- Generic Categorisation:
    - **Local Explanations:** Explains based on only an instance (e.g. why a document is relevant to a particular query?).
    - **Global Explanations:** Explains in terms of a retrieval model.

- IR Specific Categorisation:
    - ***Point-Wise Explanation:*** Explains a query-document pair.
    - ***Pair-Wise Explanation:*** Explains a pair of documents with respect to a query.
    - ***ListWise Explanation:*** Explains the ranked list corresponding to a query.

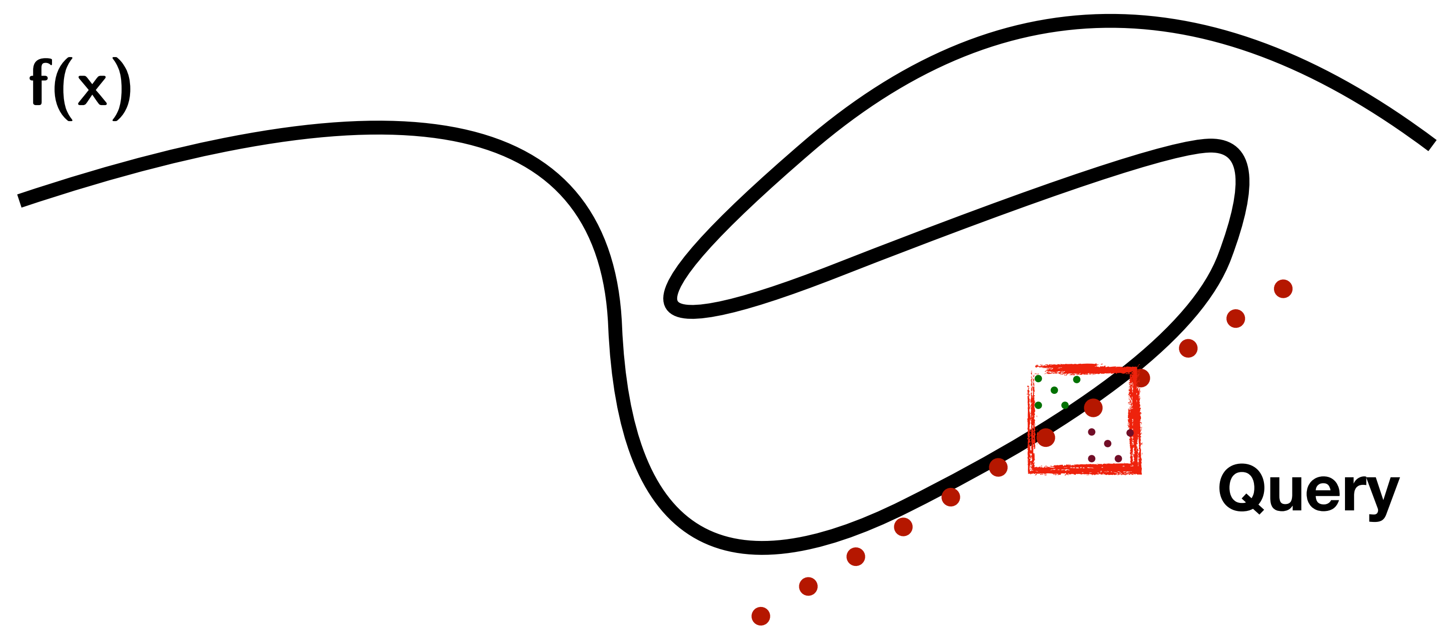
Simple vs Accuracy



- Global approximation using a simpler model and simple feature space is hard to achieve

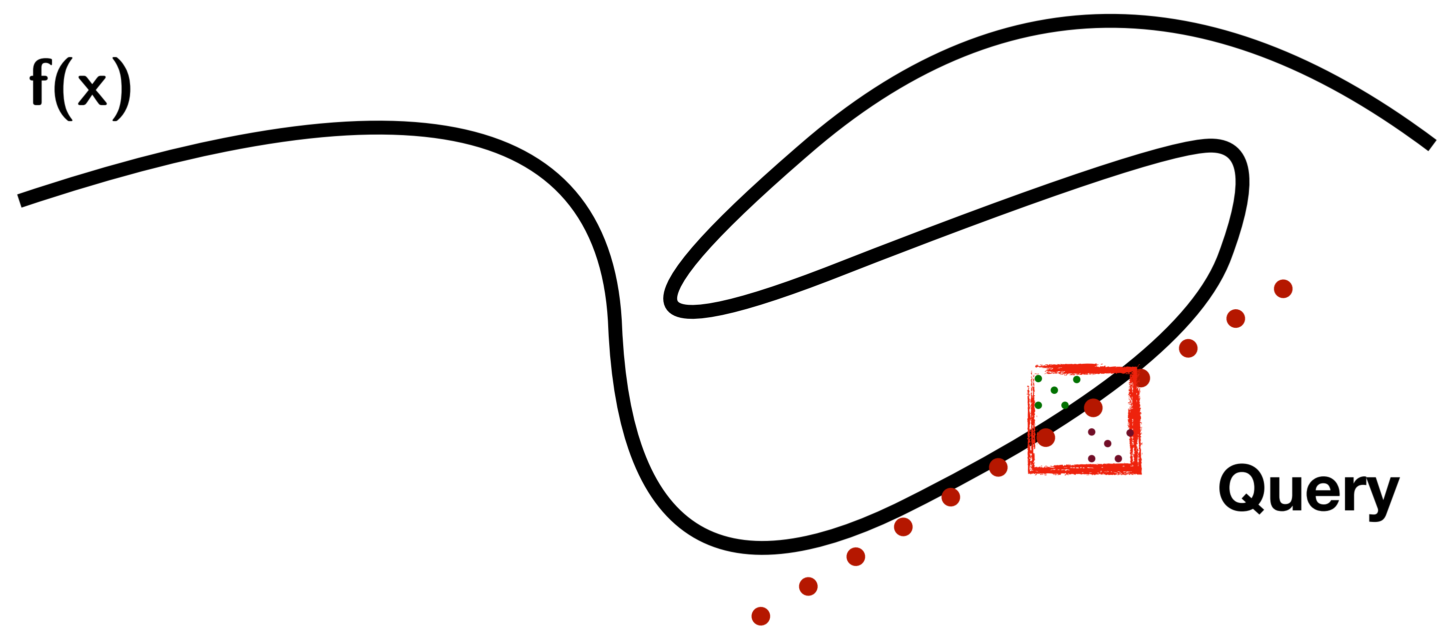
- Local approximations are possible

Local Interpretability



- Given a query instance, sample a local training dataset by querying the black box model
- Fit a simpler (proxy) model to the local dataset
- Example: LIME

LIME in a nutshell



- **Step 1:** Collect a local dataset in the epsilon neighborhood around each query instance
  - Note that the labels come from the original classifier **f(x)**
- **Step 2:** Train a simple classifier to fit the local dataset

LIRME: Adapting LIME to Rankings

- A point-wise local explanation approach for *text rankers*

- **Step 1:** Collect a local dataset in the epsilon neighborhood around each query instance

How do we create (small) perturbations to the original **text** document to create a local sample?

[Verma & Ganguly, SIGIR 19]

LIRME: Adapting LIME to Rankings

- A point-wise local explanation approach for *text rankers*

- **Step 1:** Collect a local dataset in the epsilon neighborhood around each query instance

How do we create (small) perturbations to the original **text** document to create a local sample?

- **Step 2:** Train a simple classifier to fit the local dataset

What is the simple classifier ? How do we interpret the results of the fit ?

[Verma & Ganguly, SIGIR 19]

Document Perturbations

Health hazards | Search

Doctors say fatty food is hazardous for a healthy lifestyle

~~Doctors~~ say fatty food is ~~hazardous~~ for a ~~healthy~~ lifestyle

Doctors ~~say~~ fatty food is hazardous for a healthy ~~lifestyle~~

Document Perturbations

Sample terms to be added or removed to a document

Doctors say fatty food is hazardous for a healthy lifestyle

Uniform Sampling Sample terms with a uniform likelihood (with replacement).

Biased Sampling sampling probability of a term proportional to **Tf-Idf**

Masked Sampling: Segment a document D into **D/k** chunks. Each subsample can comprise a set of chunks

LIRME :Objective Function

$\rho(D, D'_i)$: Measures the distance between D and D'_i

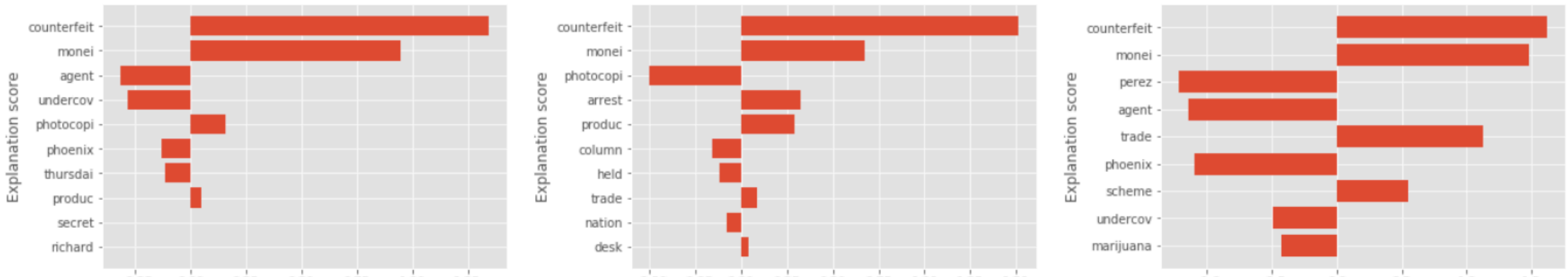
θ_j is a p dimensional vector showing the importance of a term t in $S(D, Q)$.

$$L(D, Q, \sigma, \theta) = \sum_{i=1}^{M} \rho(D, D'_i)(S(D, Q) - \sum_{j=1}^{p} \theta_j w(t_j, D'_i))^2$$

Retrieval score of a document with respect to a query.

D'_i is the sampled version of D.

Example Explanation of LIRME



(a) Uniform sampling **(b) Tf-idf sampling** **(c) Masked samples ($v = 0.1$, $k = 5$)**

Figure 4: Visualization of explanation vectors $\hat{\Theta}(Q, D)$ estimated for a sample (relevant) document 'LA071389-0111' (D) and query (Q) 'counterfeiting money' (TREC-8 id 425). The Y-axis shows explanation terms, while the X-axis plots their weights.

Evaluation Approaches Used in LIRME

- **Explanation Consistency:** Choice of samples around the pivot document **D** should not result in considerable differences in the predicted explanation vector.
  - Computes *correlation* between predicted and ground truth ranking of terms
- **Explanation Correctness:** Computes similarity between *explanation vector terms* $\theta(Q, D)$ and *relevant terms* $R(Q)$

LIRME: Explanation from an IR Practitioner's POV

- Pointwise and global explanation approach
- **Explanation units** — term frequency, document length, document frequency, semantic similarity
- Provides a framework to explain both
  - within a ranking model and
  - between different retrieval models

[Verma & Ganguly, SIGIR 20]

Global Feature Importance

- For each retrieval model and for each query train a regression classifier based on the fundamental features

- Choose randomly k number of queries for a particular model

- Contribution of each feature is the average weights learned across K queries

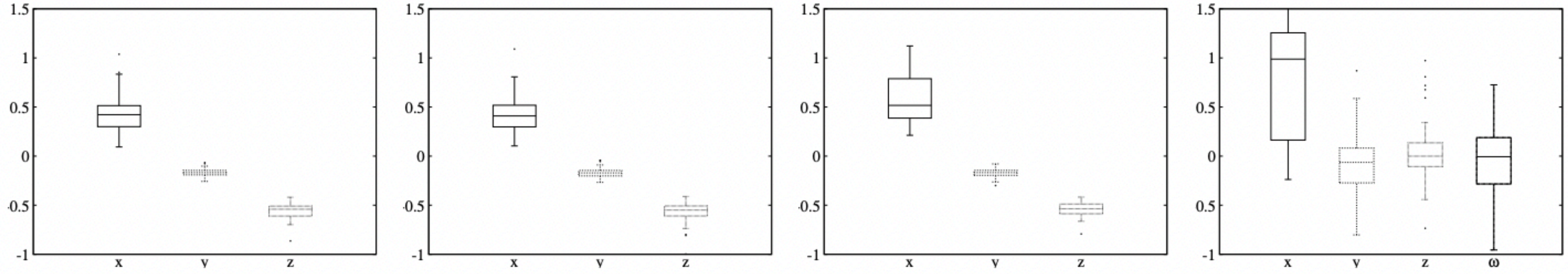


Figure 1: Box-plot of parameter vectors θ for BM25, LM-JM, LM-Dir and DRMM (in order from left-right).

Explanation Within a Ranking Model

- Why does a model M retrieve a document D_1 at rank r_1 and D_2 at r_2 ($r_2 > r_1$ without loss of generality) for a query Q (Pair-Wise Explanation)?
- Compute the contribution of a feature in the retrieval score computation.
- Compute relative Contribution Difference between a pair of documents.
- If Fidelity score and importance of the feature have same signs, that acts as a possible explanation

Explanation Across Ranking Models

- Why does a model M_1 retrieve a document D at position r_1, whereas model M_2 retrieves D at r_2 for a query Q?
- $\xi(M_1, M_2) = \Delta_s(D, M_1, M_2) \cdot \Delta(M_1, M_2)$,
- $\Delta(M_1, M_2) = \vec{\theta}(M_1, Q) - \vec{\theta}(M_2, Q)$ measures the relative importance difference between the feature importance across different retrieval models.
- $\Delta_s(D, M_1, M_2)$ measures the relative drop in score with respect to the top most document.
- If $\xi_x > 0$ that acts as a plausible explanation.

Listwise explanations

We have to explain an already trained model **f(Q)**

$$C = \{D_1, \ldots, D_N\} \xrightarrow{\textbf{f(Q)}} D_1 > D_2 > \ldots > D_k$$

$\updownarrow$ **Rank Correlation**

$$C = \{D_1, \ldots, D_N\} \xrightarrow{\textbf{E(Q')}} D_1 > D_2 > \ldots > D_k$$

- **Explanation:** A set of terms that is a super set of Q
  - Q’ = Q U {w1 , w2, …} where wi are explanation terms
- **Proxy Model:** A simple and easy to understand model

Local Interpretability for Rankings

$$C = \{D_1, \ldots, D_N\} \xrightarrow{\textbf{f(Q)}} D_1 > D_2 > \ldots > D_k$$

Query Q

Query Q
+

Explanation Terms

f(Q) $\longrightarrow$ **E(Q')**

Trained Model

Proxy Model

Trained Model	Proxy Model
d_1	d_1
d_3	d_3
d_5	d_5
d_2	d_2
d_4	d_4

Rank Correlation

Selecting Candidate Terms

Health hazards | Search

Doctors say fatty food is hazardous for a healthy lifestyle 0.93

~~Doctors~~ say fatty food is ~~hazardous~~ for a ~~healthy~~ lifestyle 0.03

Doctors ~~say~~ fatty food is hazardous for a healthy ~~lifestyle~~ 0.92

Candidate Terms

{doctor, hazardous, healthy}

Preserving Rank Correlation

Health hazards doctor

d1

d2

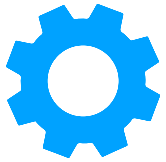
d3

d4

d5

Preference Pairs

Candidate Terms	d1 > d2	d1 > d3	d2 > d3	d1 > d4	d2 > d5
le					
handle					
doctor		0.38			
invert					
medicin					

How much does "doctor" prefer d1 over d3 using 

Preference Coverage Problem

$$\max \sum_{0 \le j < m} [\![y_j > 0]\!]$$

$$s.t.$$

$$s_i \in \{0, 1\}, 0 \le i < n$$

$$y_j = \sum_{0 \le i \le n} s_i . P_{i,j} . w_{i,j}$$

NP-Hard: Generalization of budgeted max. weighted coverage

Solution: Greedy heuristic and ILP works well in practice

Evaluating Explanations

Query

Ground Truth ?

Query
+
Expansion Terms

Explanation
Accuracy

f(Q)
Trained Model

d_1
d_3
d_5
d_2
d_4

0.78

Fidelity

E(Q')
Proxy Model

d_1
d_3
d_5
d_2
d_4

Now improved

Query

Ground Truth ?

Query
+
Expansion Terms

Explanation
Accuracy

f(Q)
Trained Model

E(Q')
Proxy Model

d_1
d_3
d_5
d_2
d_4

0.78

Fidelity

> 0.90

d_1
d_3
d_5
d_2
d_4

Anecdotal Results

Query	Intent Explanation
alexian brothers hospital (DRMM)	patient course war person sister leader alliance
alexian brothers hospital (DESM)	medication treating memory nurses father physical doctors
afghanistan flag (DRMM)	US official inscription time transit dave november
afghanistan flag (DESM)	symbol nation flagpole hoist general banner flagstaff
fidel castro (DRMM)	havana domestic cuba invest intestine real medical
fidel castro (DESM)	cuban havana dictator communist president raul gonzalez
how to find the mean (DRMM)	x statistics plus know
how to find the mean (DESM)	actually say want meant

Recent Results

Multiple Explainers: Rankings with different aspects

Query: Bobcat

Explainers	Explanation
Term Matching	charlotte, north, sales, 2008
Position Awareness	basketball, north, states, learn
Semantic Similarity	felidae, carnivorous, boko extinction, deserts, iucn
Multiplex (Multiple Explainers)	felidae, carnivorous, boko extinction, deserts, gvwr, north

Lyu & Anand, ECIR'23

Free-Text Explanations: Overview

- Free Text explanations methods aim to generate explanations using natural language.
- Typical free-text explanations are not more than a few sentences long, and sometimes even limited to a few words.
- Approaches for text ranking models focus either on interpreting the query intent as understood by a ranking model or on producing a short text summary to explain why an individual document or a list of documents is relevant.

Query Intent Explanation

- **Input:** Query, Set of Relevant Documents, Set of Irrelevant documents.
- **Output:** Intent Description which precisely interprets the search intent that can help distinguish the relevant documents from the irrelevant documents.
- Exploits an Encoder Decoder Architecture.

Query Intent Descriptor Architecture

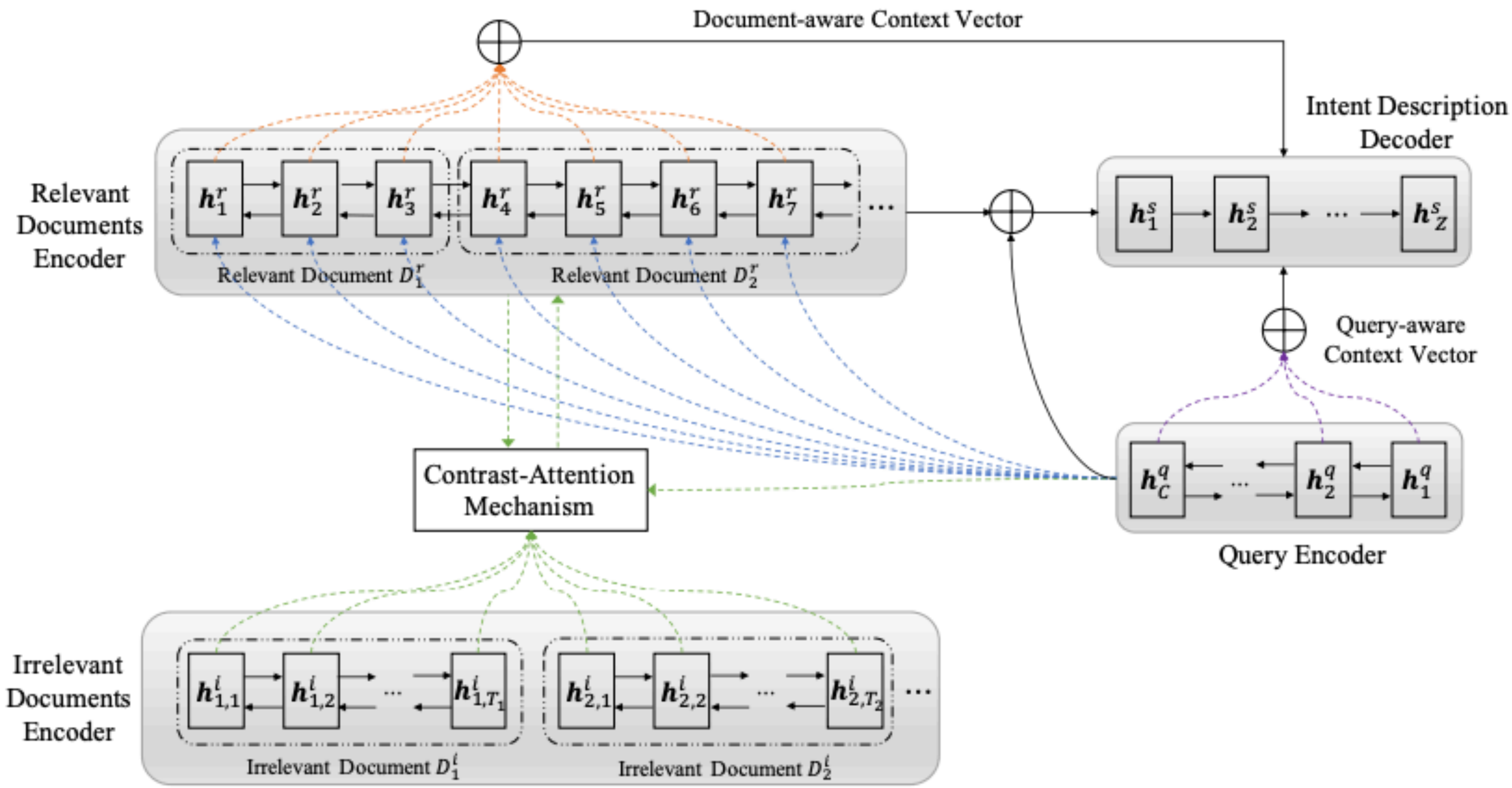


Figure 1: The overall architecture of contrastive generation model (CtrsGen).

Architecture of Intent Descriptor

Example Intent Description

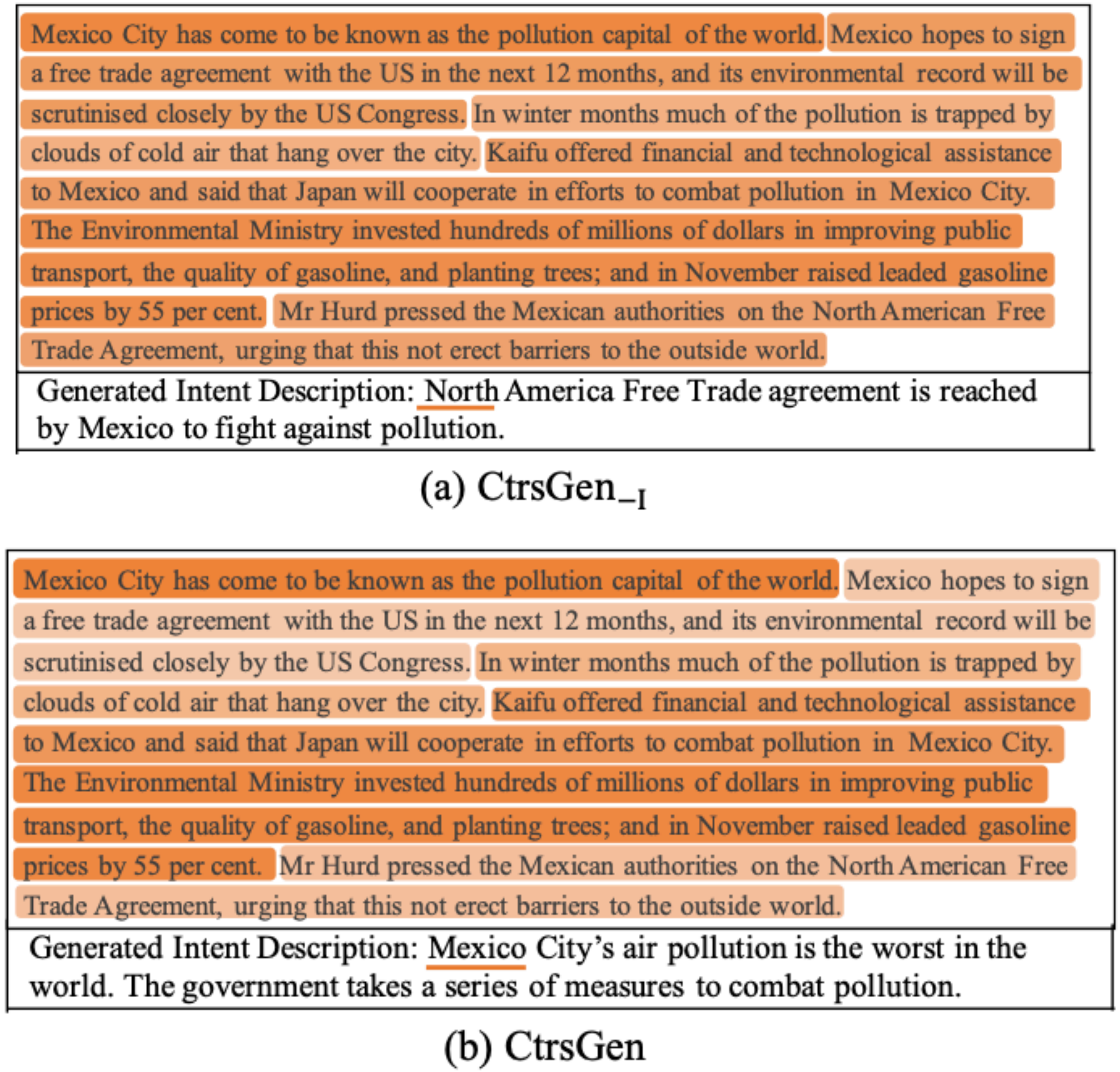


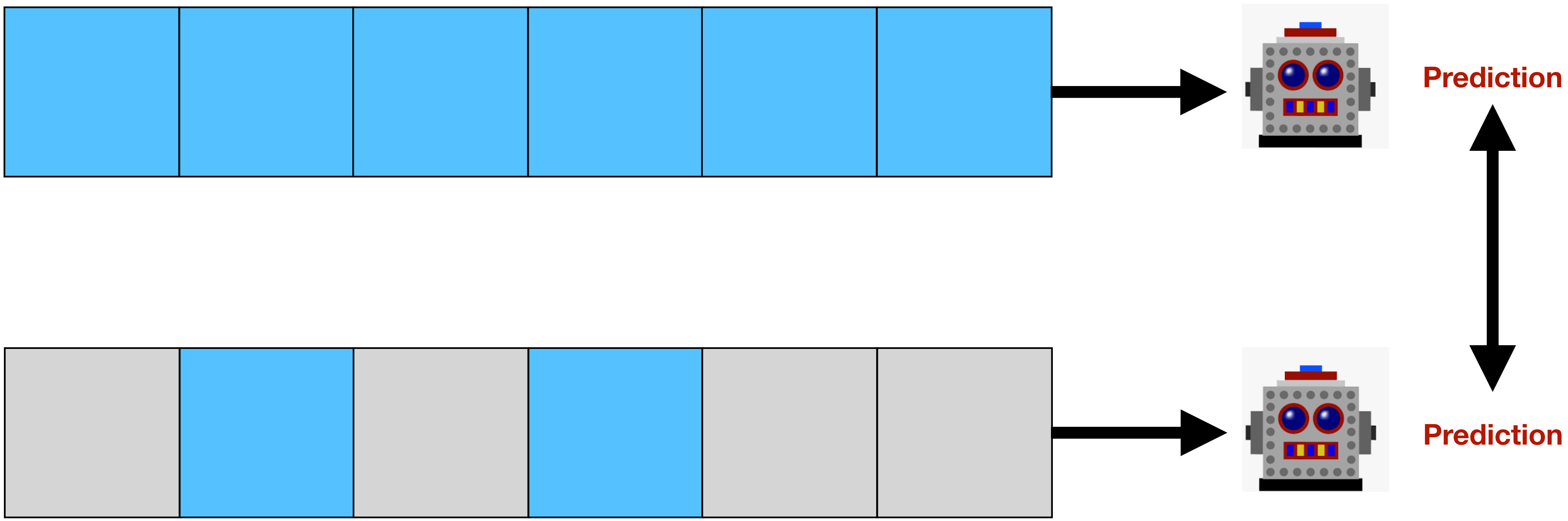
Figure 3: (a) and (b) is the heatmap of the sentence-level decoder attention weights in relevant documents for generating the first word in the description, given by *CtrsGen-$_I$* and *CtrsGen* respectively. Deeper shading denotes higher value.

Thank You

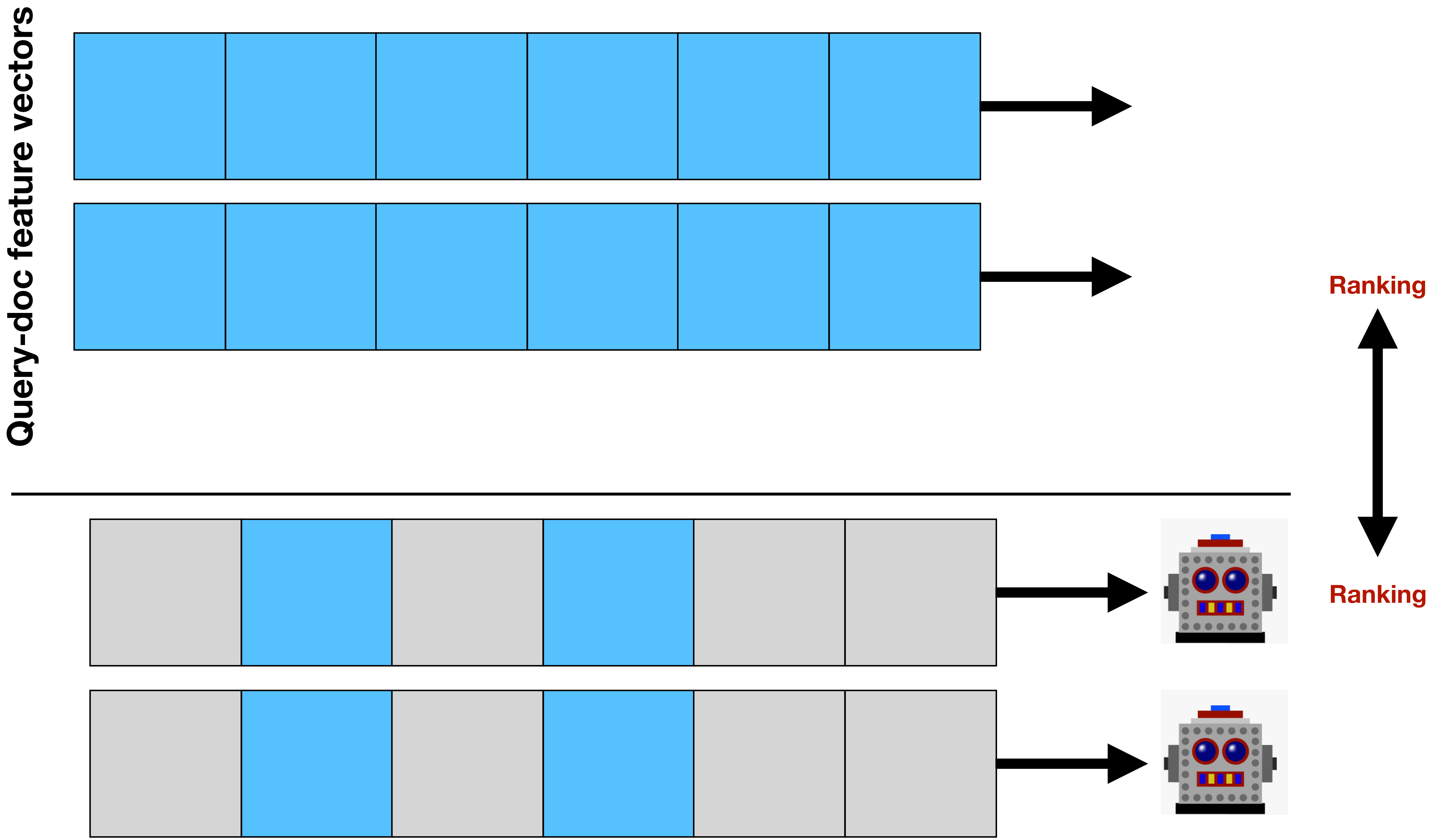
Information theoretic view

*"**Feature-based** explanations are valid if they contain most of the predictive power"*

Explanations from RDT

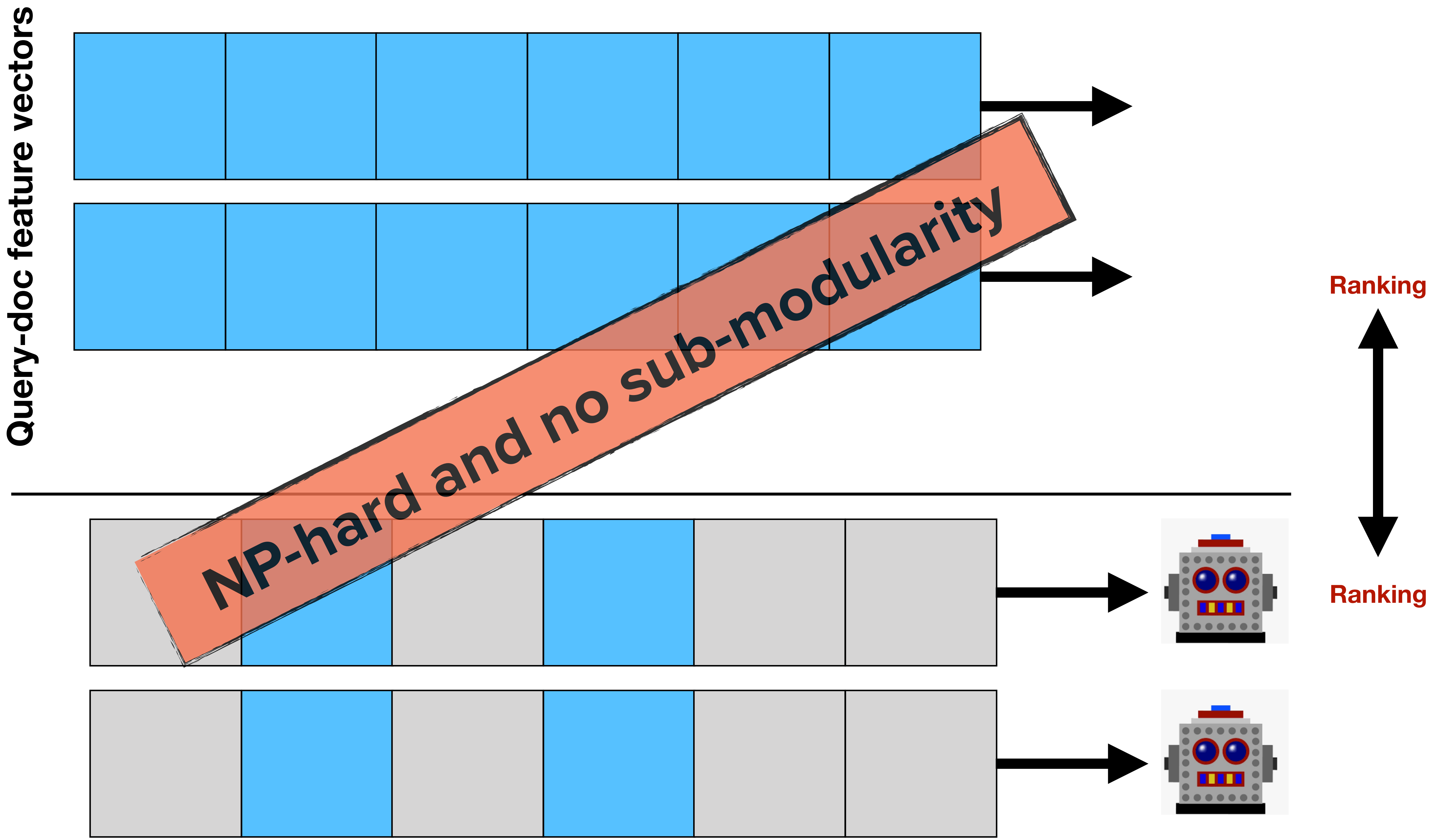


Local Ranking Explanations



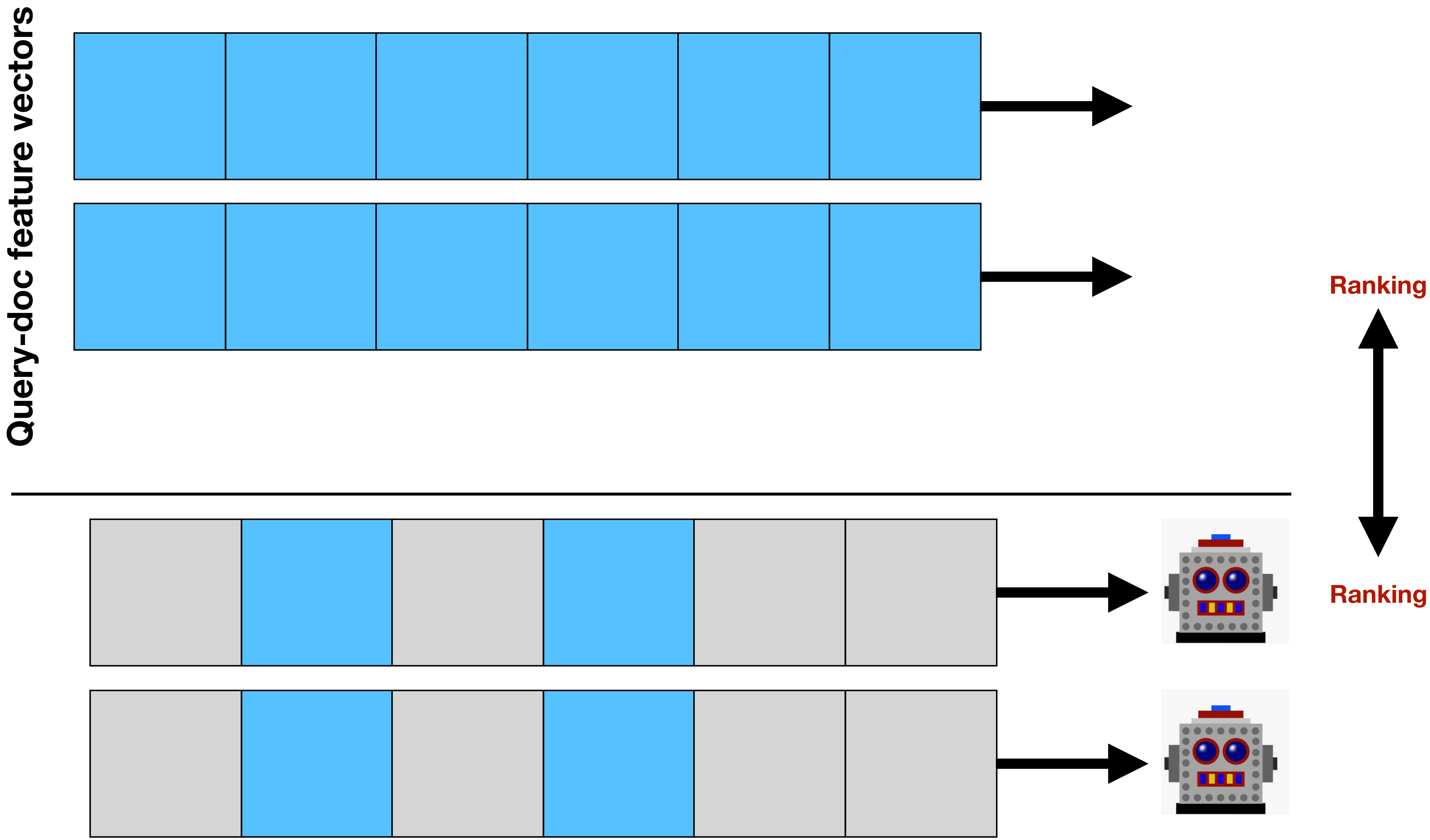
Extracting per Query Valid Explanations for Blackbox Learning-to-Rank Models, Singh et al., *ICTIR 21.*

Local Ranking



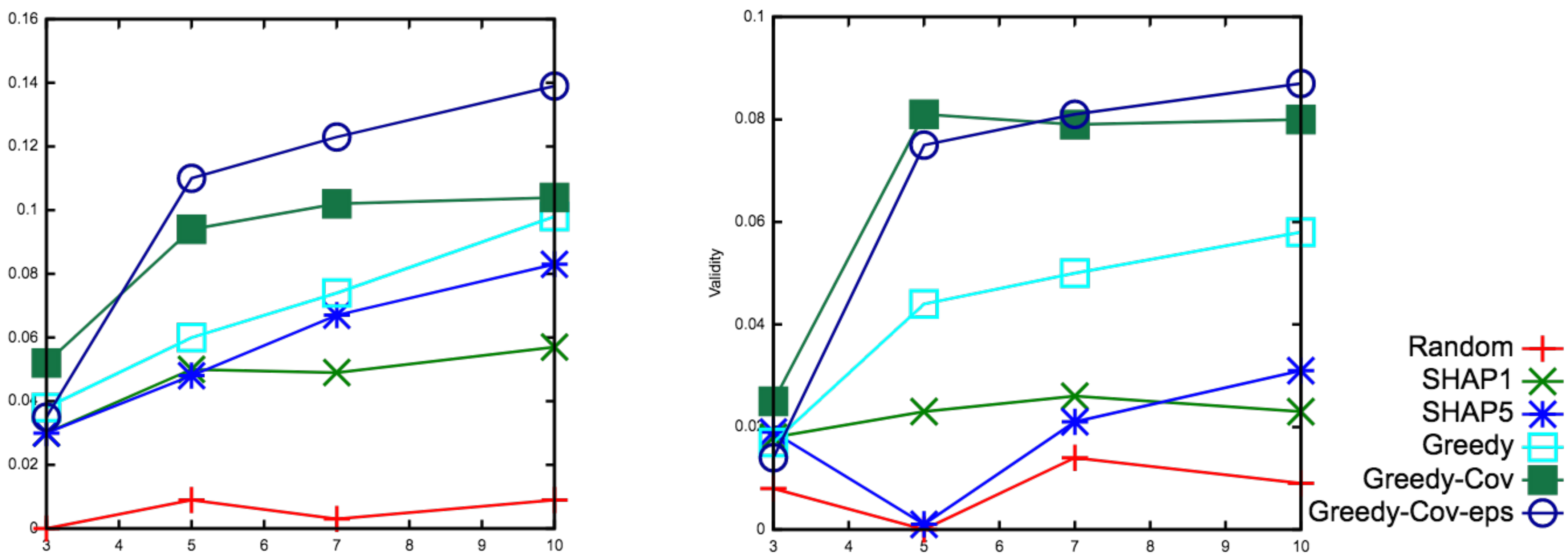
Extracting per Query Valid Explanations for Blackbox Learning-to-Rank Models, Singh et al., *ICTIR 21.*

Greedy Algorithm



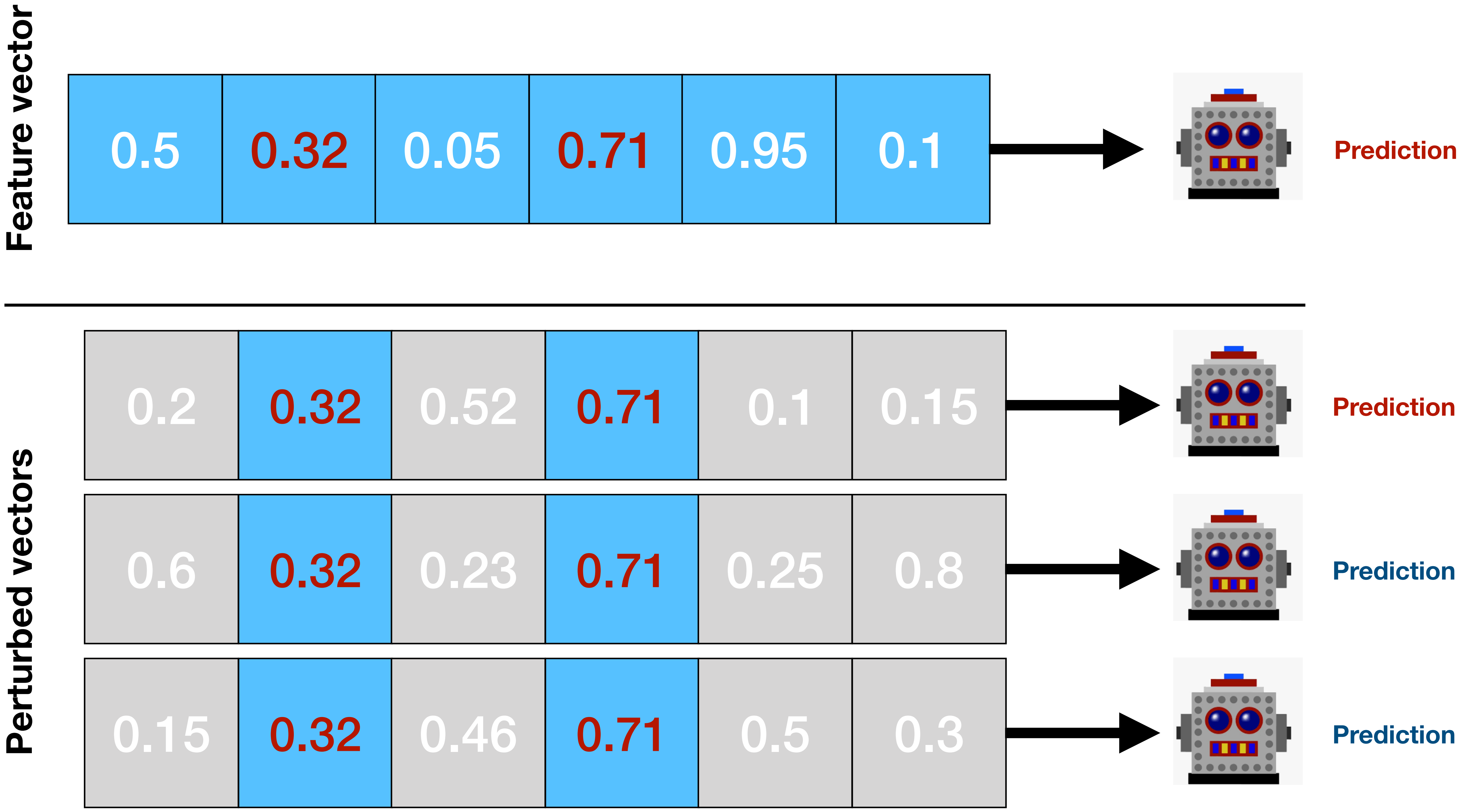
Extracting per Query Valid Explanations for Blackbox Learning-to-Rank Models, Singh et al., *ICTIR 21.*

Result



(a) Pairwise Ranker (b) Listwise Ranker

Stable Explanations



Stable Explanations

Problem: Choose subset of explanation features that result in majority of reconstructions being aligned to the original prediction