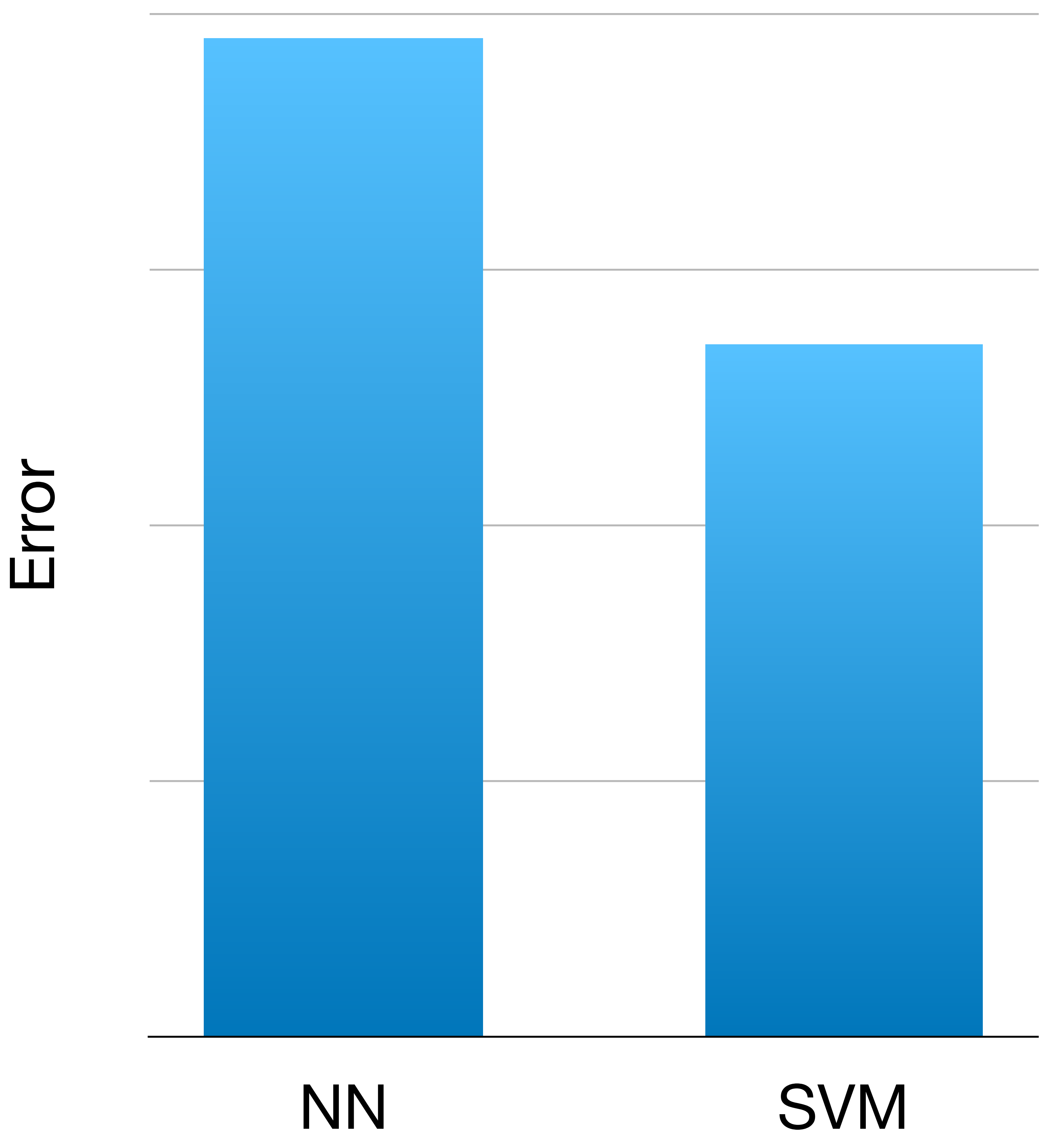
Deriving Machine Attention from Human Rationales

Yujia Bao[1]

Shiyu Chang[2], Mo Yu[2], Regina Barzilay[1]

[1]Computer Science and Artificial Intelligence Lab, MIT
[2]MIT-IBM Watson AI Lab, IBM Research

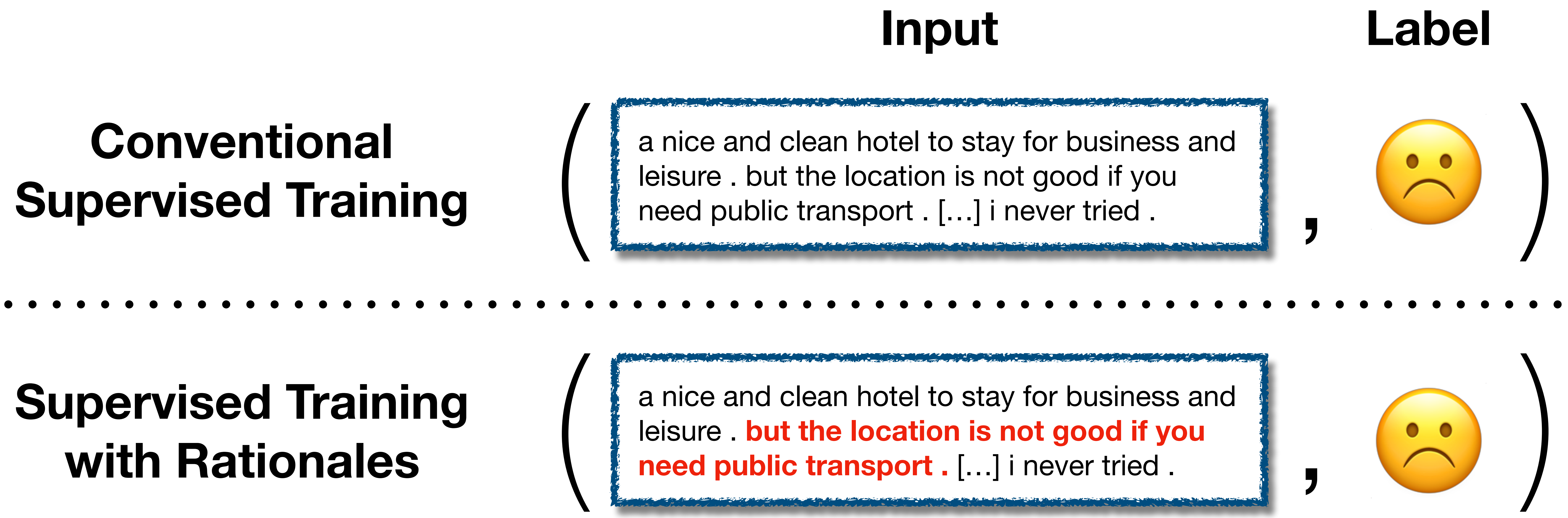
Neural Networks in Low-resource Scenario



Training data: **200** instances

Can NN do better on small training sets?

Human Rationales can Help



	Input		Label
Conventional Supervised Training	(a nice and clean hotel to stay for business and leisure . but the location is not good if you need public transport . […] i never tried .	,	☹)
Supervised Training with Rationales	(a nice and clean hotel to stay for business and leisure . **but the location is not good if you need public transport .** […] i never tried .	,	☹)

- **Rationales are useful for training SVMs [1]**
- **Limited benefits for neural models [2]**

1. Zaidan et al., Using annotator rationales to improve machine learning for text categorization, NAACL 2007.
2. Zhang et al., Rationale-augmented convolutional neural networks for text classification, EMNLP 2016.

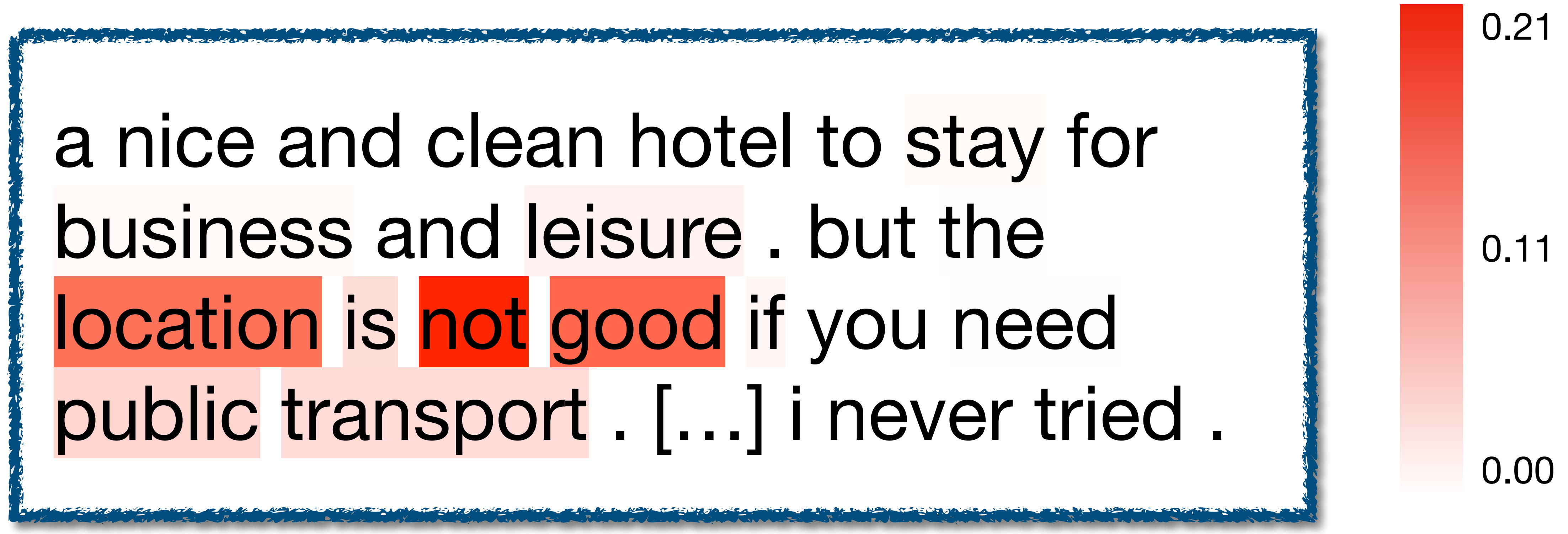
Rationales and Attention are Closely Linked

Rationales

a nice and clean hotel to stay for business and leisure . **but the location is not good if you need public transport .** […] i never tried .

Task: *hotel location*

Attention (#data 14K)



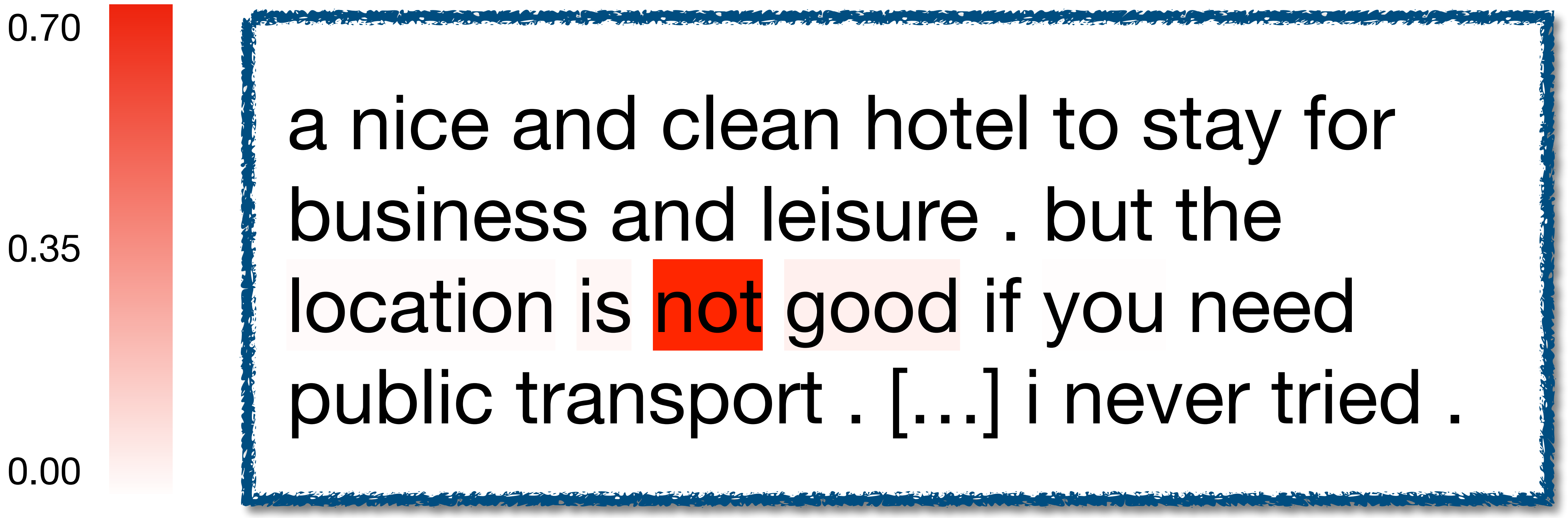
a nice and clean hotel to stay for business and leisure . but the location is not good if you need public transport . […] i never tried .

Task: *hotel location*

Both highlight important words from the input.

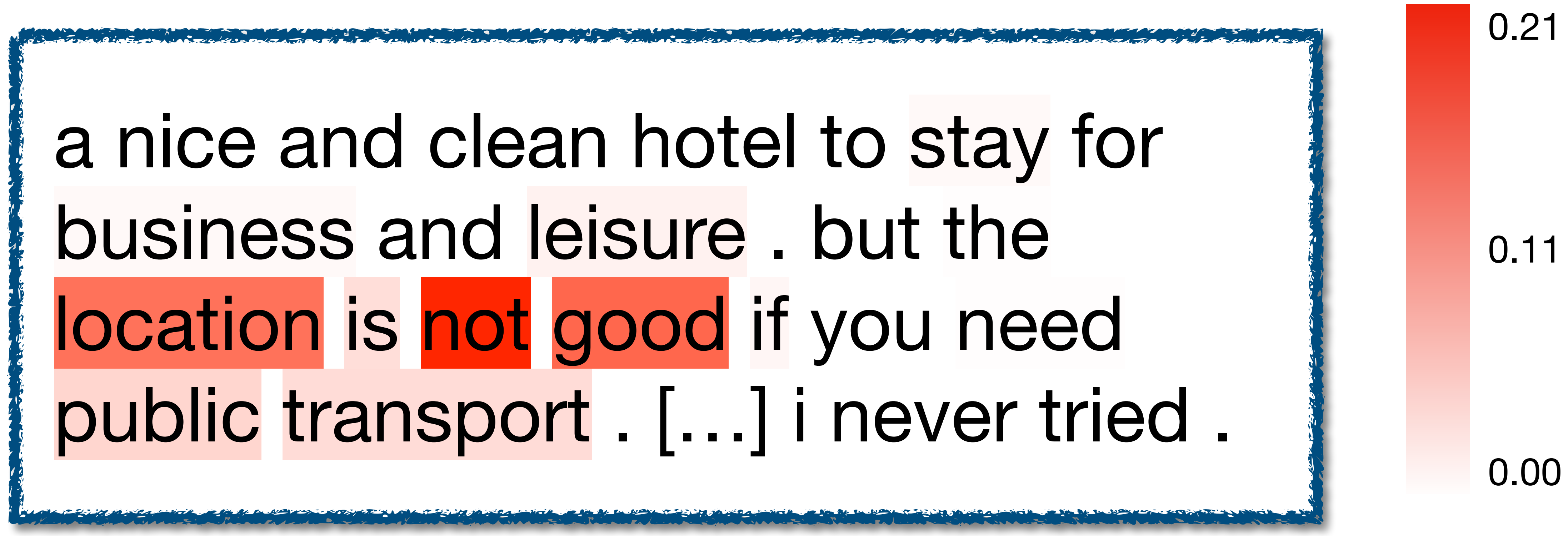
Attention in Low-resource Scenario

Attention (#data 200)



a nice and clean hotel to stay for business and leisure . but the location is not good if you need public transport . […] i never tried .

Attention (#data 14K)

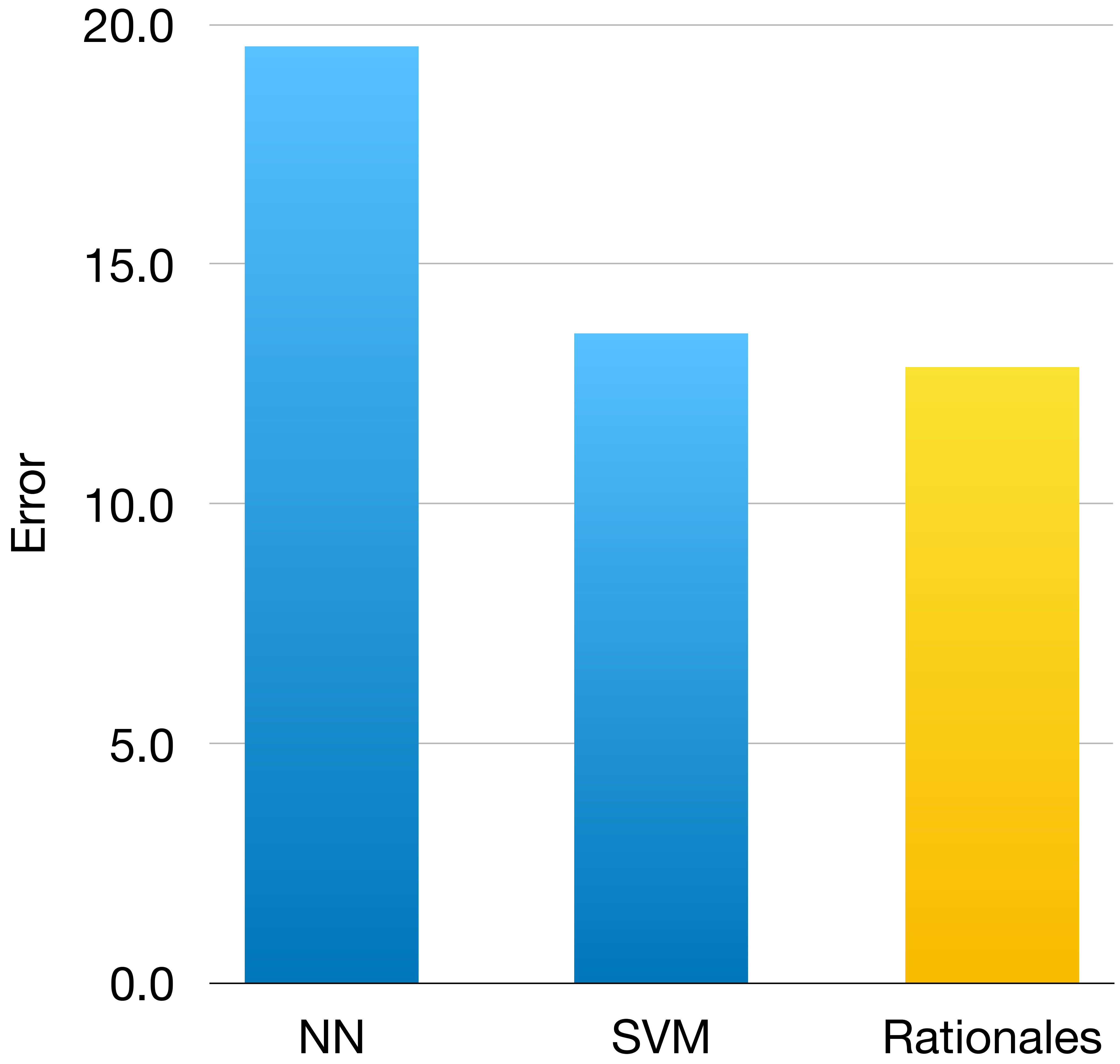


a nice and clean hotel to stay for business and leisure . but the location is not good if you need public transport . […] i never tried .

Difficult to learn where to focus

Can we use human rationales to directly supervise attention?

Human Rationales as Attention Supervision: A Naive Approach



Training objective

- Prediction error (as before)
- Distance between learned attention and human rationales.

Can we do better?

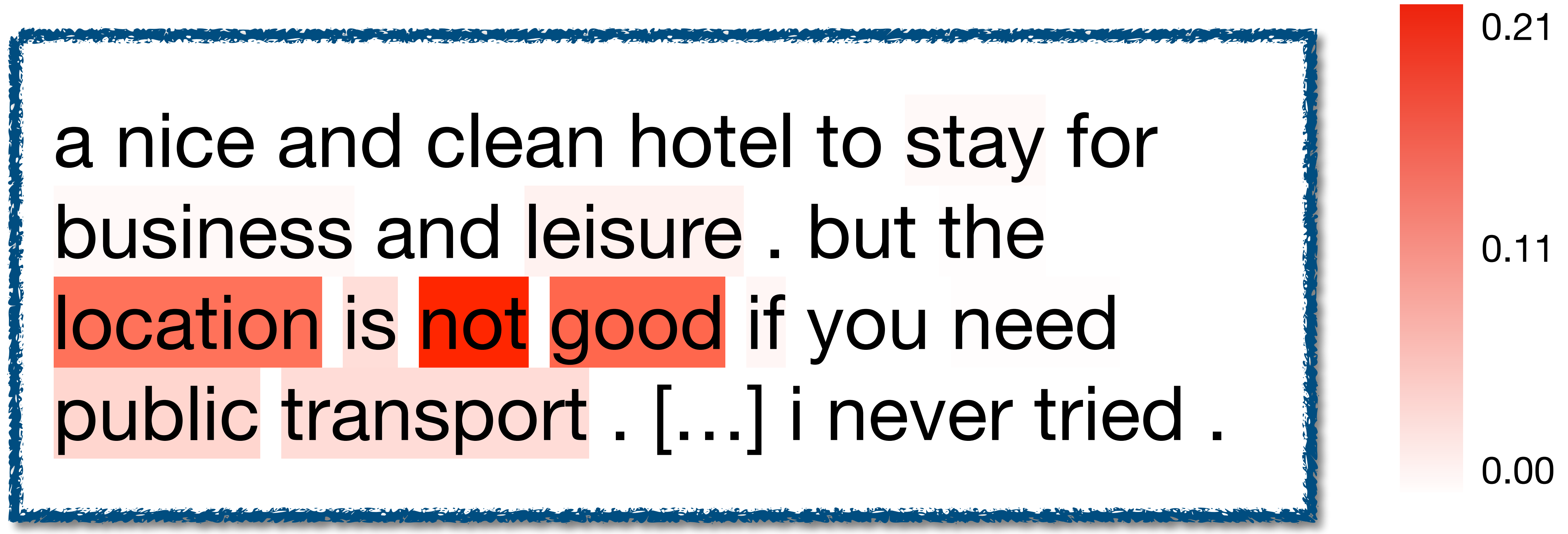
Difference between Rationales and Attention

Rationales

a nice and clean hotel to stay for business and leisure . **but the location is not good if you need public transport .** […] i never tried .

Task: *hotel location*

Attention (#data 14K)

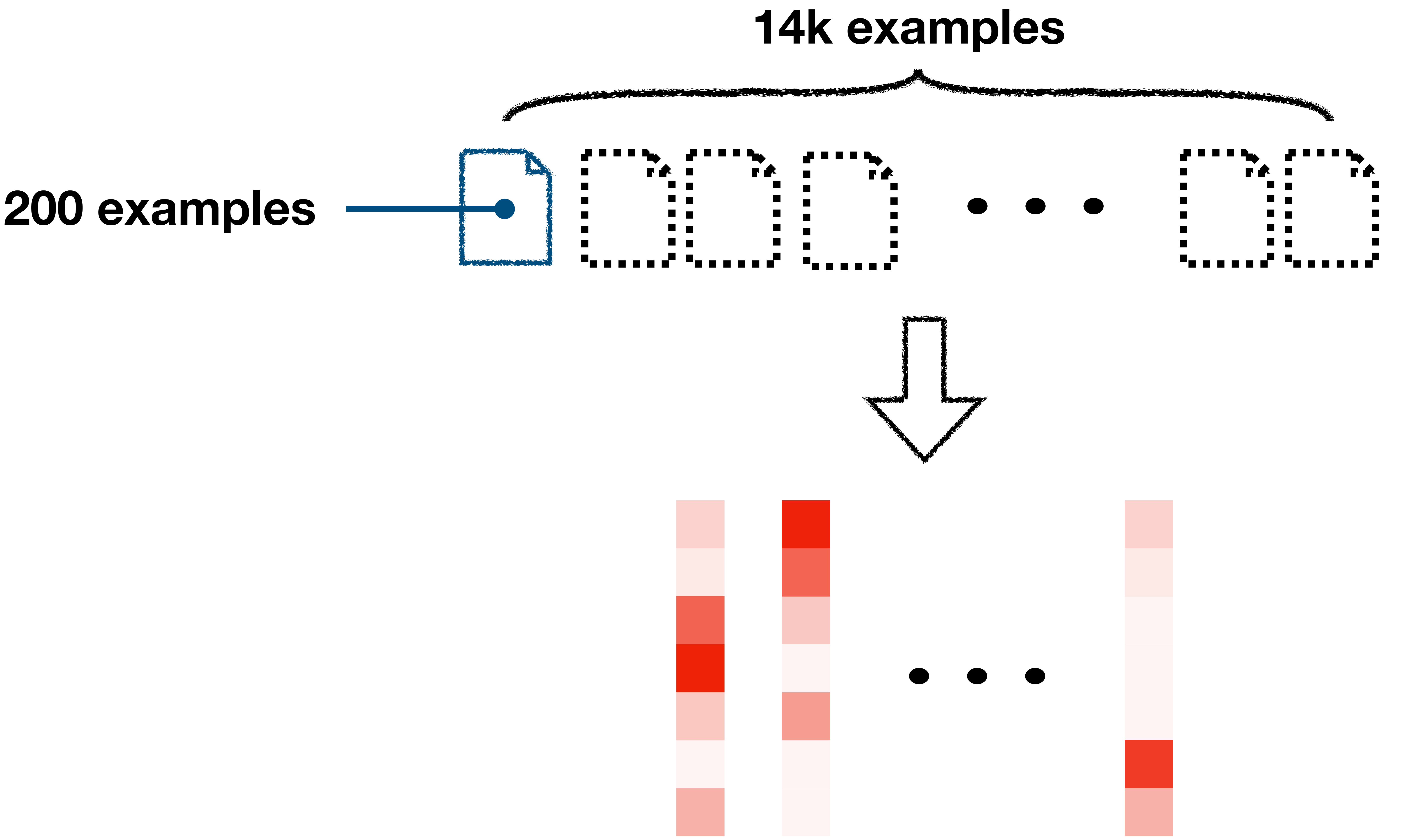


a nice and clean hotel to stay for business and leisure . but the location is not good if you need public transport . […] i never tried .

Task: *hotel location*

- Attention is a soft distribution over the input
- Attention depends on the model architecture
- Rationales are subjectively annotated

Learning with Oracle Attention

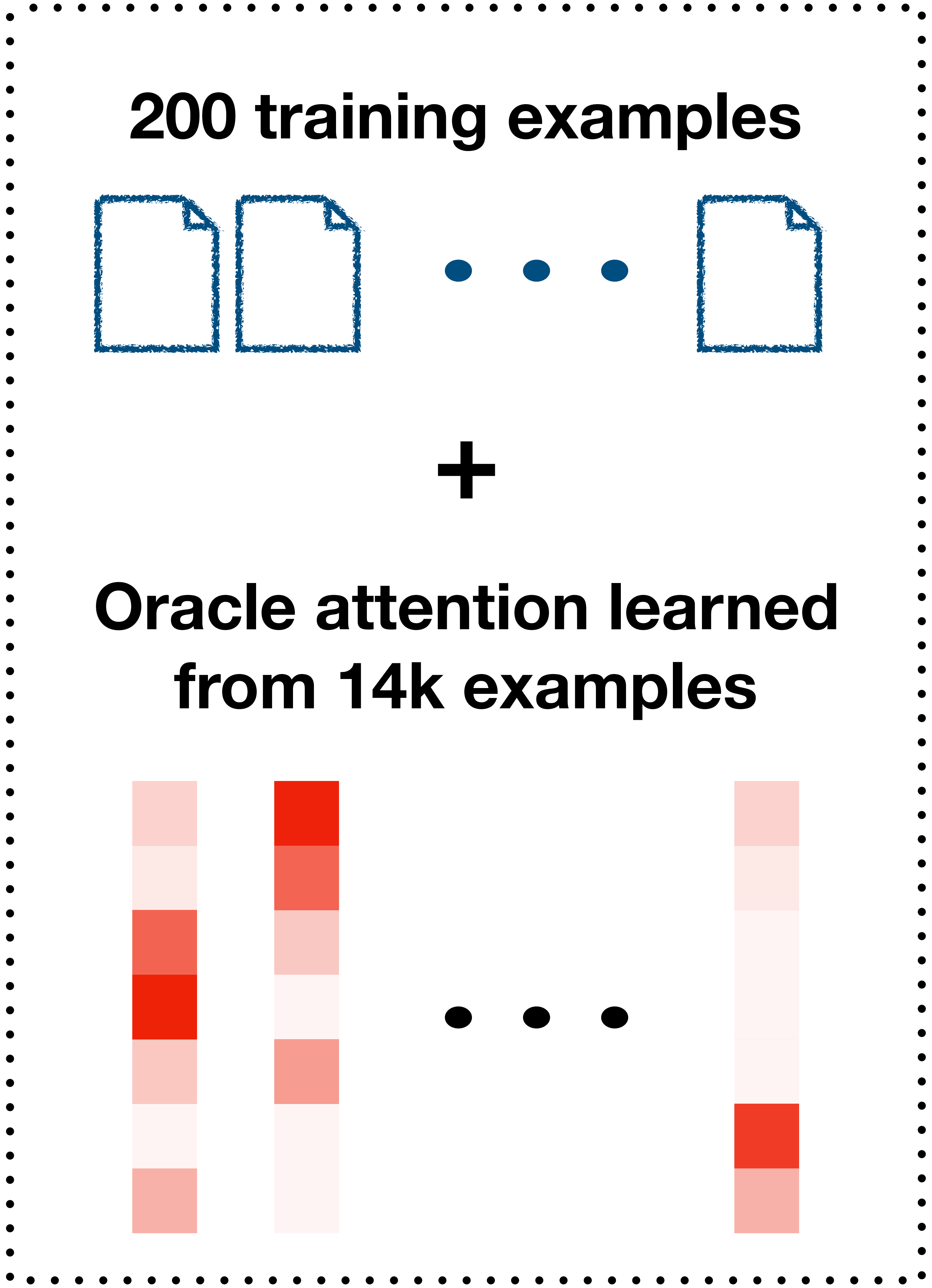


14k examples

200 examples

Oracle attention learned from 14k examples

Learning with Oracle Attention

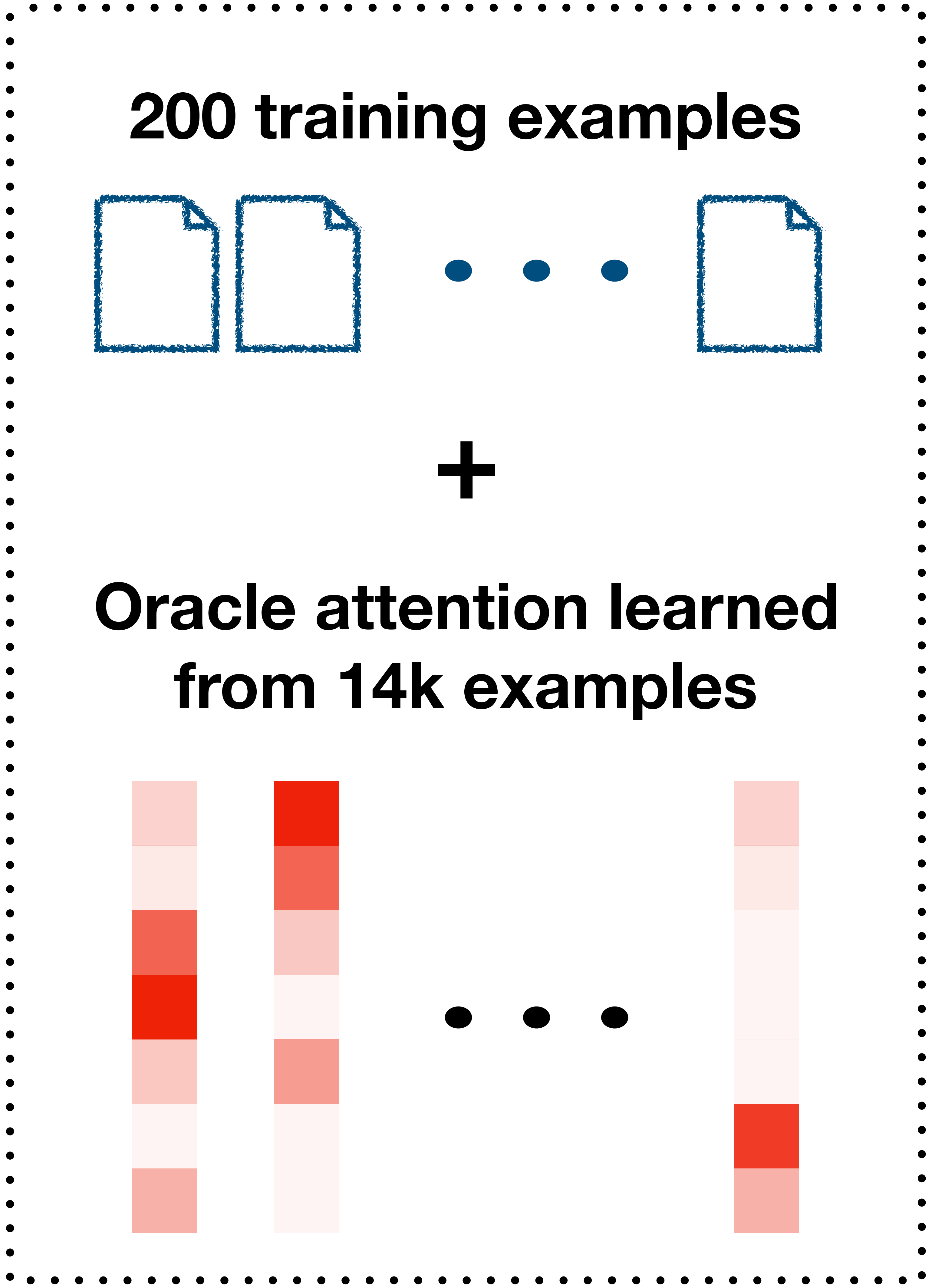


200 training examples

+

Oracle attention learned from 14k examples

Learning with Oracle Attention

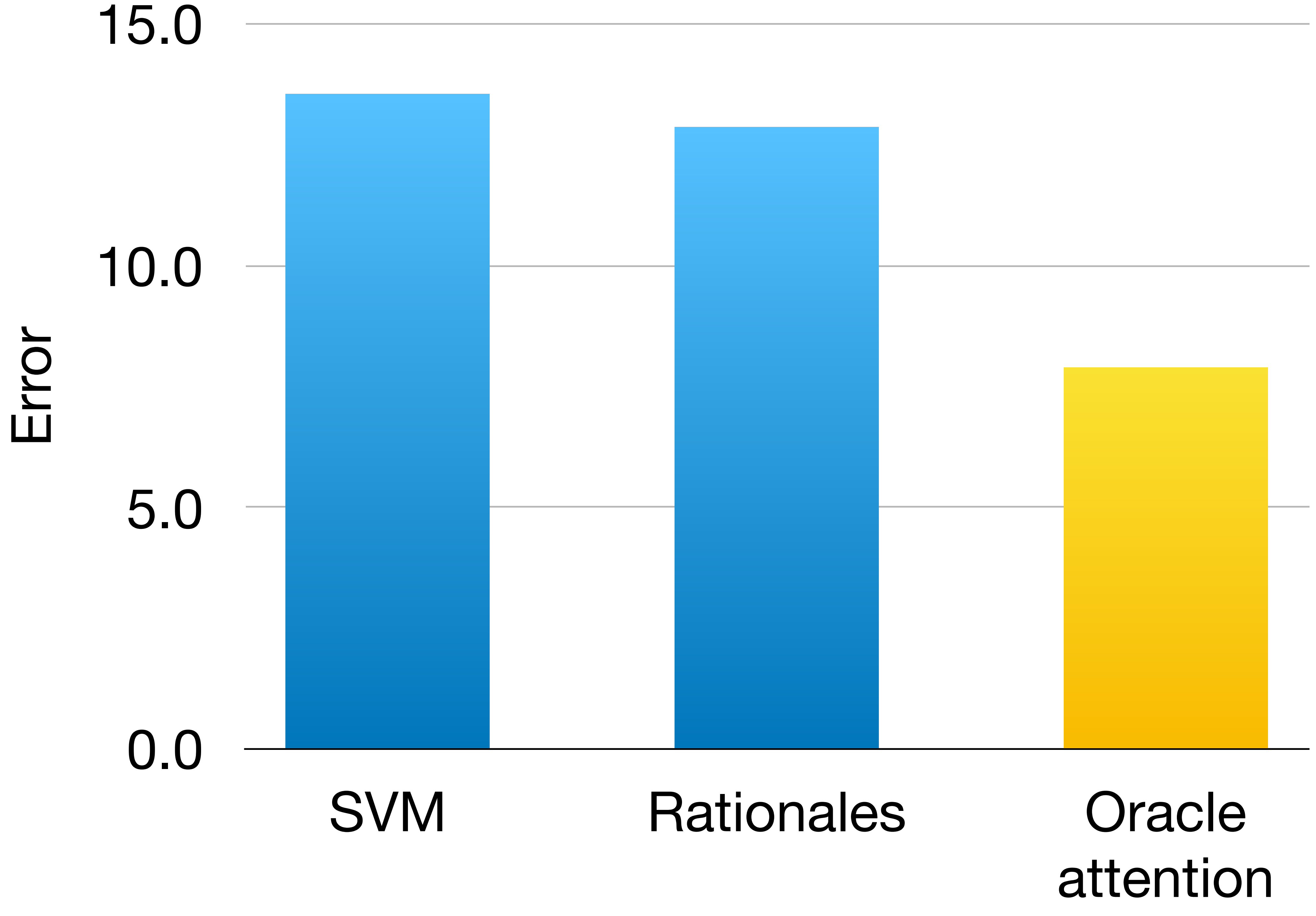


200 training examples

+

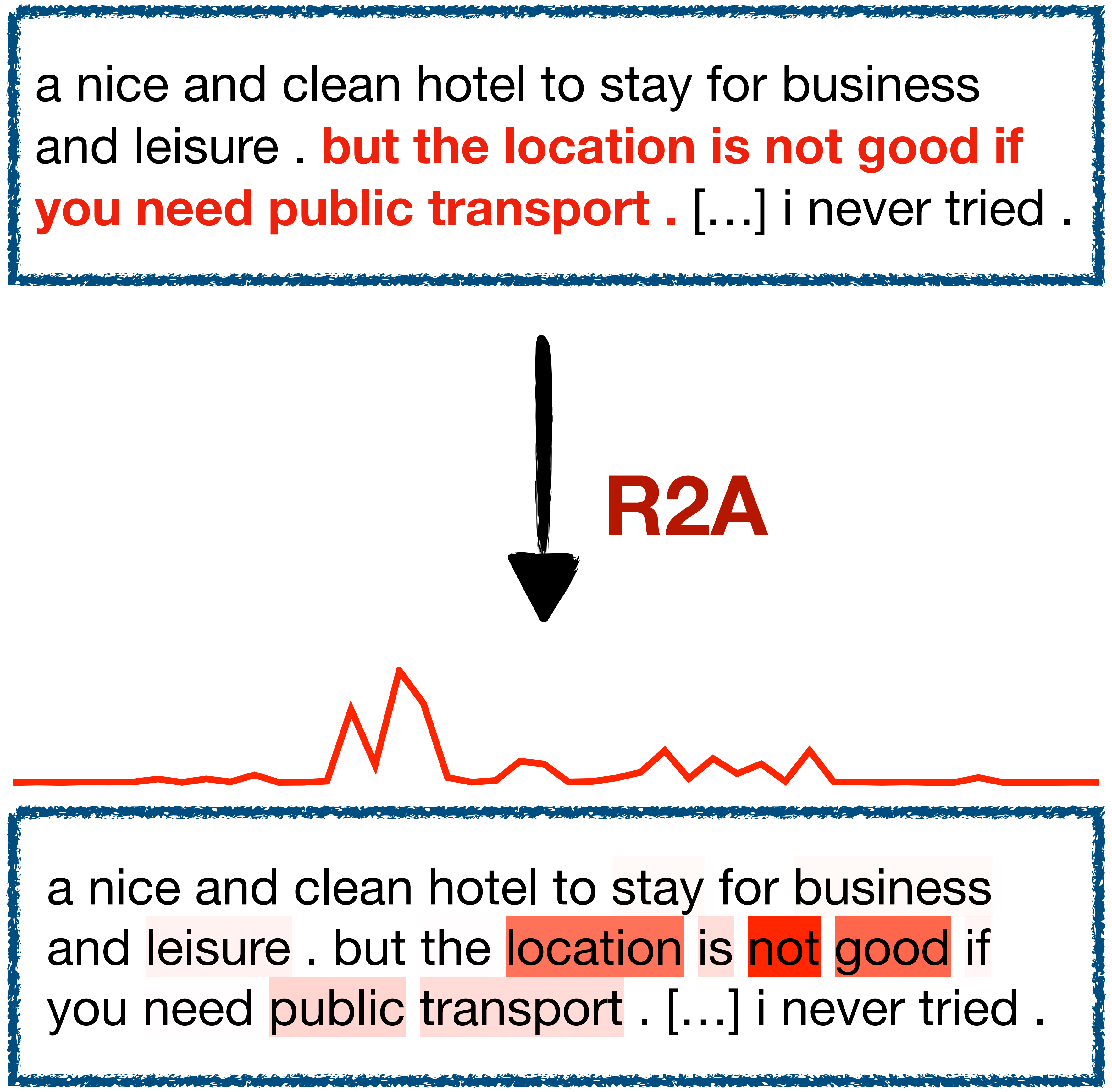
Oracle attention learned from 14k examples

38% error reduction!



Goal: translate rationales into a proxy for oracle attention.

Rationale to Attention (R2A)



a nice and clean hotel to stay for business and leisure . **but the location is not good if you need public transport .** […] i never tried .

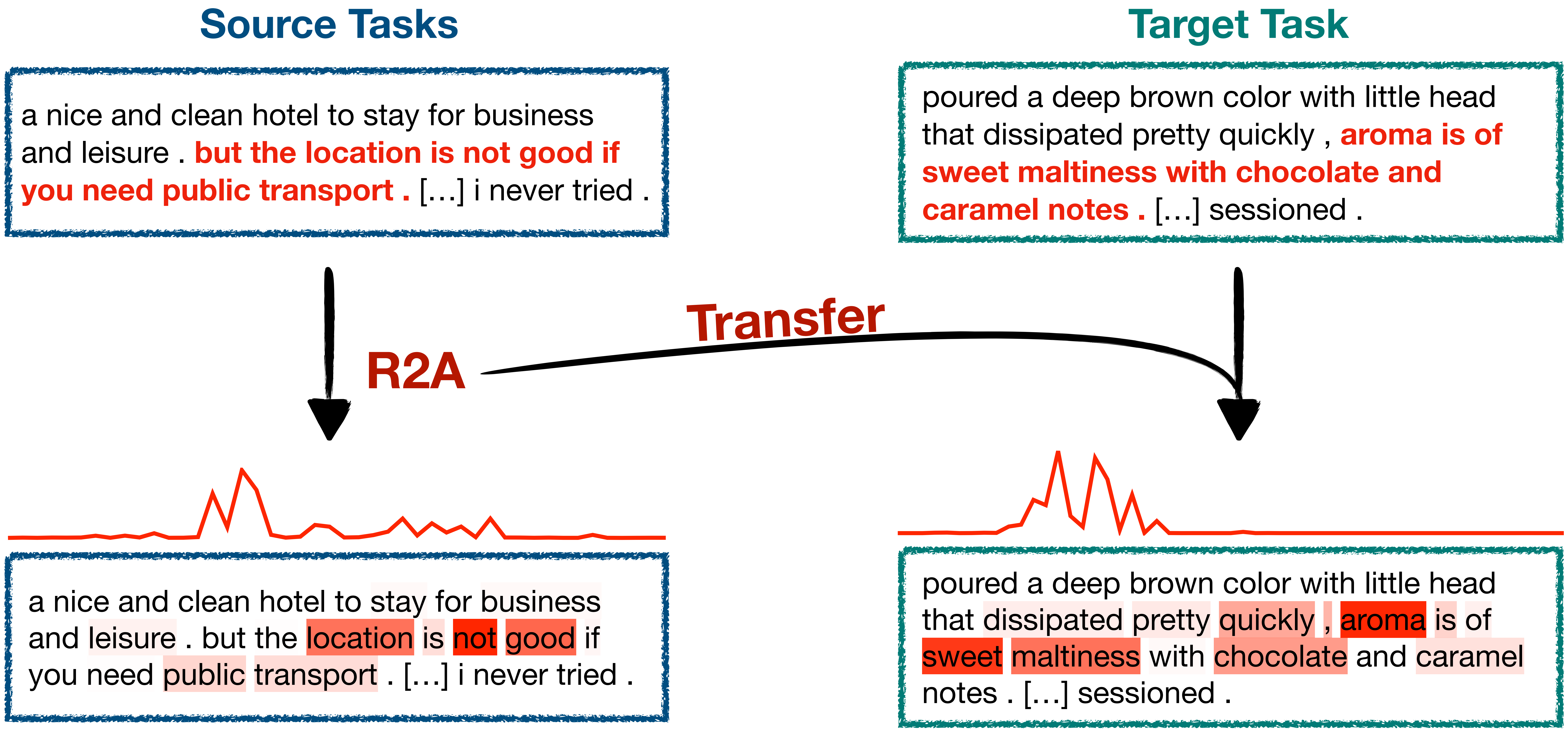
R2A

a nice and clean hotel to stay for business and leisure . but the location is not good if you need public transport . […] i never tried .

Observations:

- Attention concentrates on rationales.
- Attention highlights adjectives and nouns.
- Attention down weighs functional words

Rationale to Attention (R2A)



Source Tasks

a nice and clean hotel to stay for business and leisure . **but the location is not good if you need public transport .** […] i never tried .

Target Task

poured a deep brown color with little head that dissipated pretty quickly , **aroma is of sweet maltiness with chocolate and caramel notes .** […] sessioned .

R2A

Transfer

a nice and clean hotel to stay for business and leisure . but the location is not good if you need public transport . […] i never tried .

poured a deep brown color with little head that dissipated pretty quickly , aroma is of sweet maltiness with chocolate and caramel notes . […] sessioned .

Hypothesis: the mapping R2A is transferrable across tasks.

R2A as Attention Supervision

Step 1:

Train R2A on source tasks.

Step 2:

Use R2A to generate attention for the target task.

Step 3:

Train a target classifier with R2A-generated attention.

R2A as Attention Supervision

Step 1:

Train R2A on source tasks.

Step 2:

Use R2A to generate attention for the target task.

Step 3:

Train a target classifier with R2A-generated attention.

R2A as Attention Supervision

Step 1:

Train R2A on source tasks.

Step 2:

Use R2A to generate attention for the target task.

Step 3:

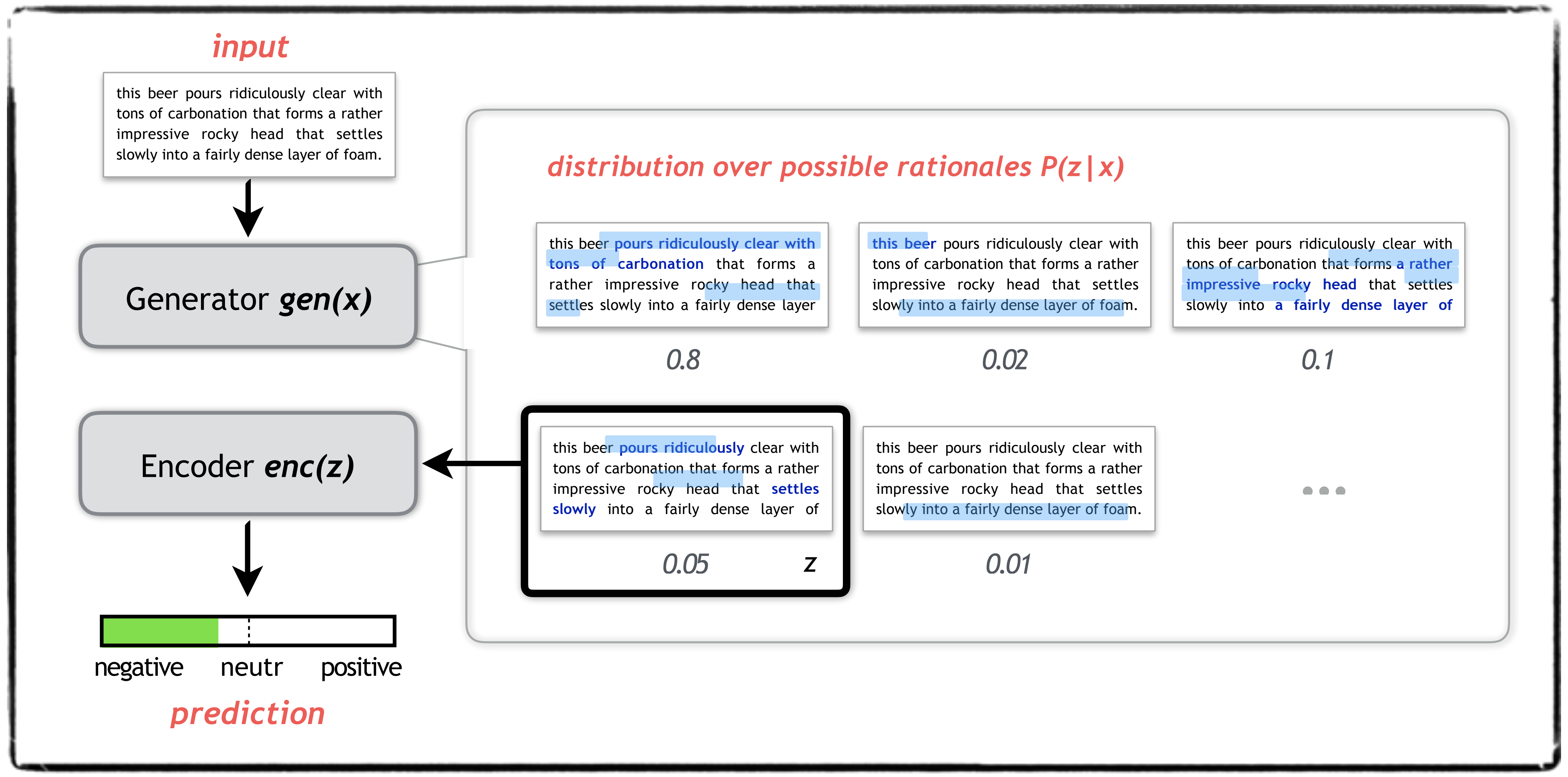
Train a target classifier with R2A-generated attention.

Where do rationales come from?

Target task: rationales are annotated by human

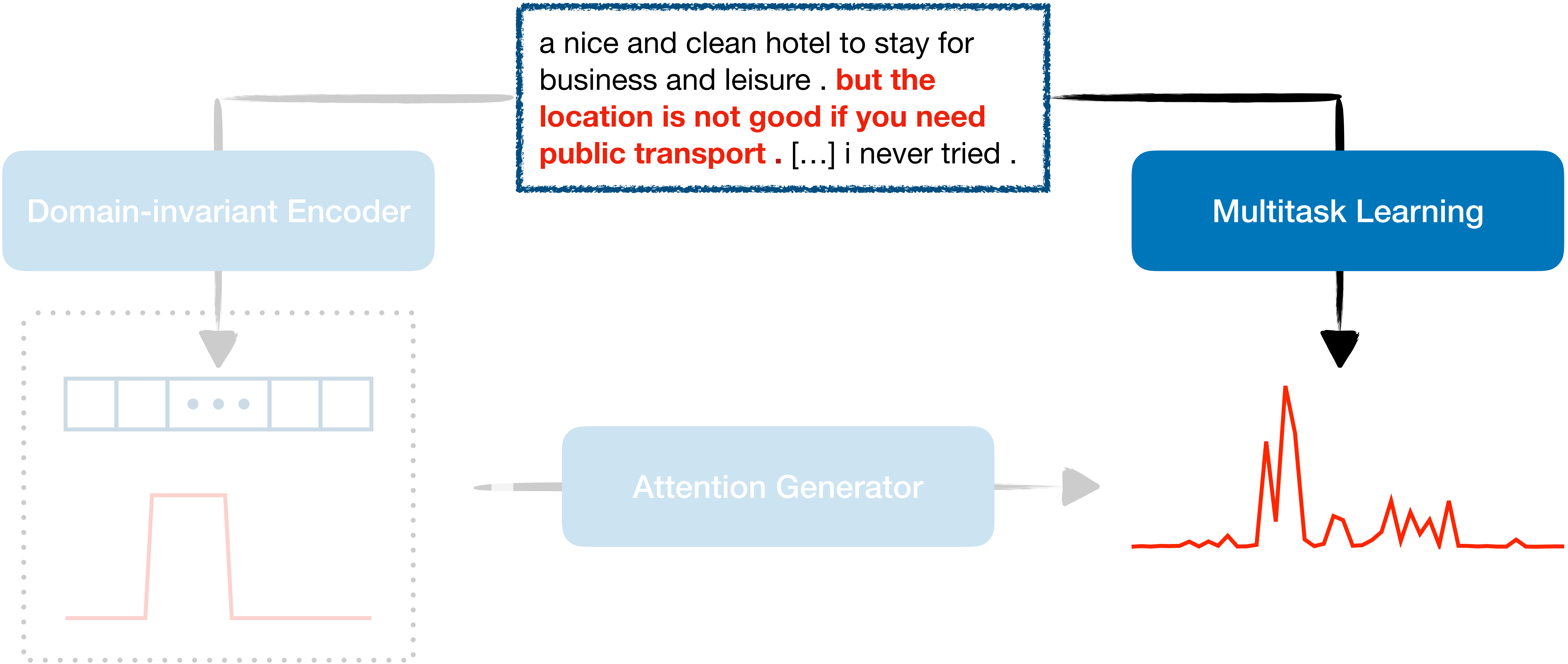
- 2x annotation cost [1]

Source tasks: rationales are generated automatically [3]


input
this beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam.
Generator gen(x)
Encoder enc(z)
distribution over possible rationales P(z|x)
0.8
0.02
0.1
0.05
z
0.01
...
negative
neutr
positive
prediction

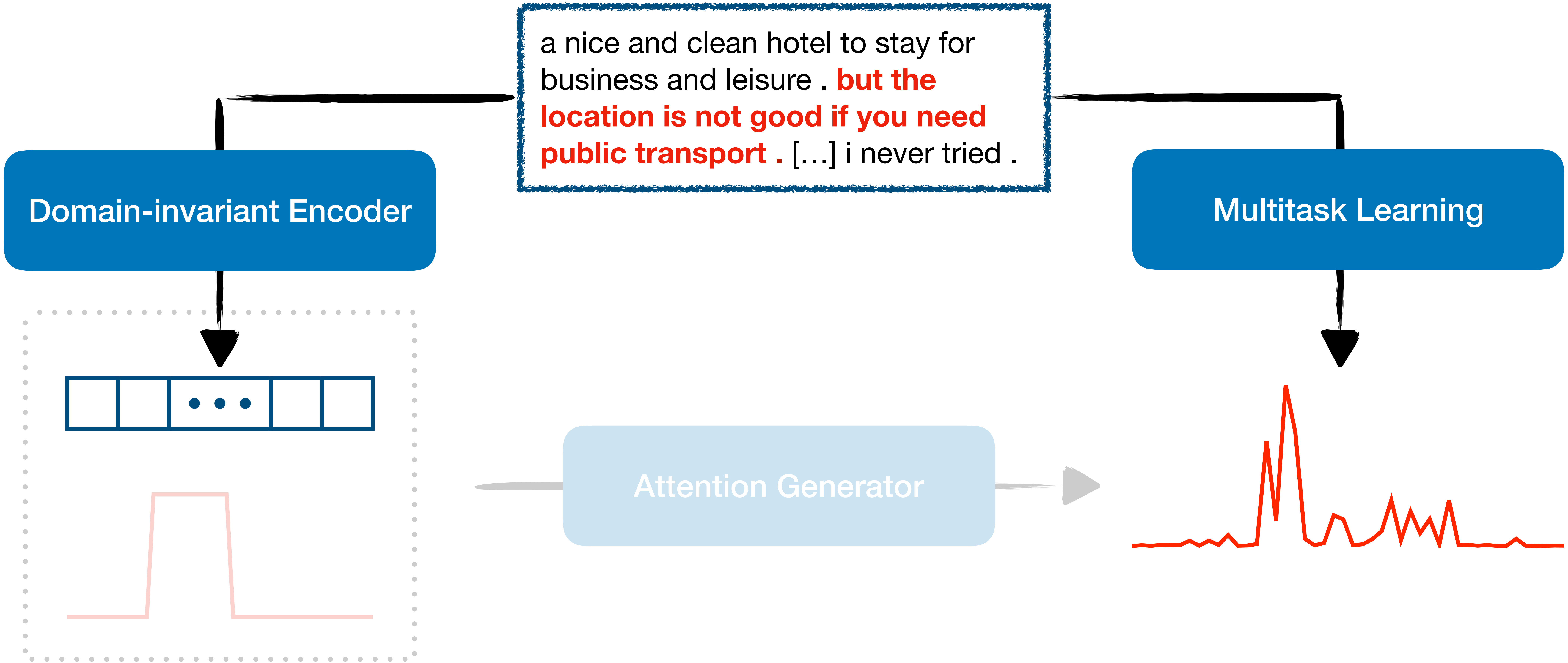
[3] Lei et al., Rationalizing neural predictions. EMNLP 2016.

R2A Training



a nice and clean hotel to stay for business and leisure . **but the location is not good if you need public transport .** […] i never tried .

R2A Training



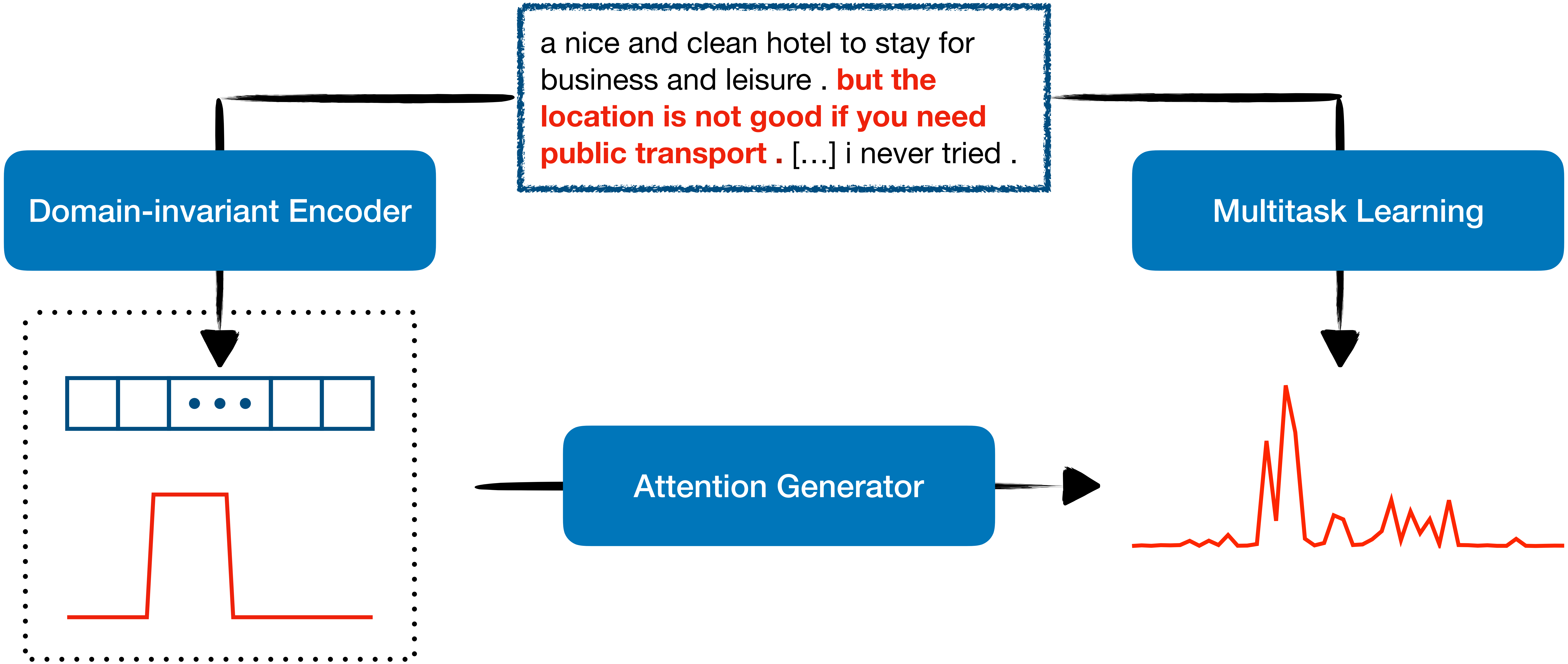
a nice and clean hotel to stay for business and leisure . **but the location is not good if you need public transport .** […] i never tried .

Domain-invariant Encoder

Multitask Learning

Attention Generator

R2A Training



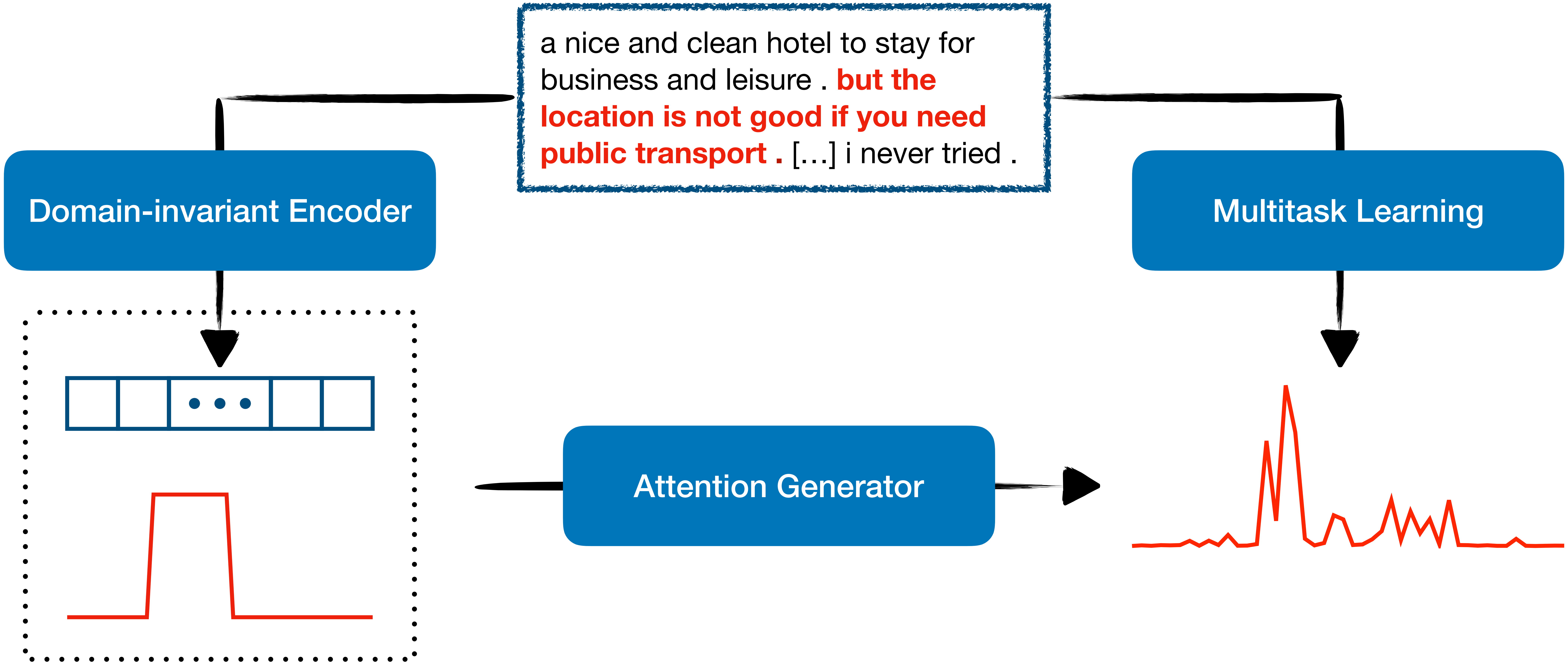
a nice and clean hotel to stay for business and leisure . **but the location is not good if you need public transport .** […] i never tried .

Domain-invariant Encoder

Multitask Learning

Attention Generator

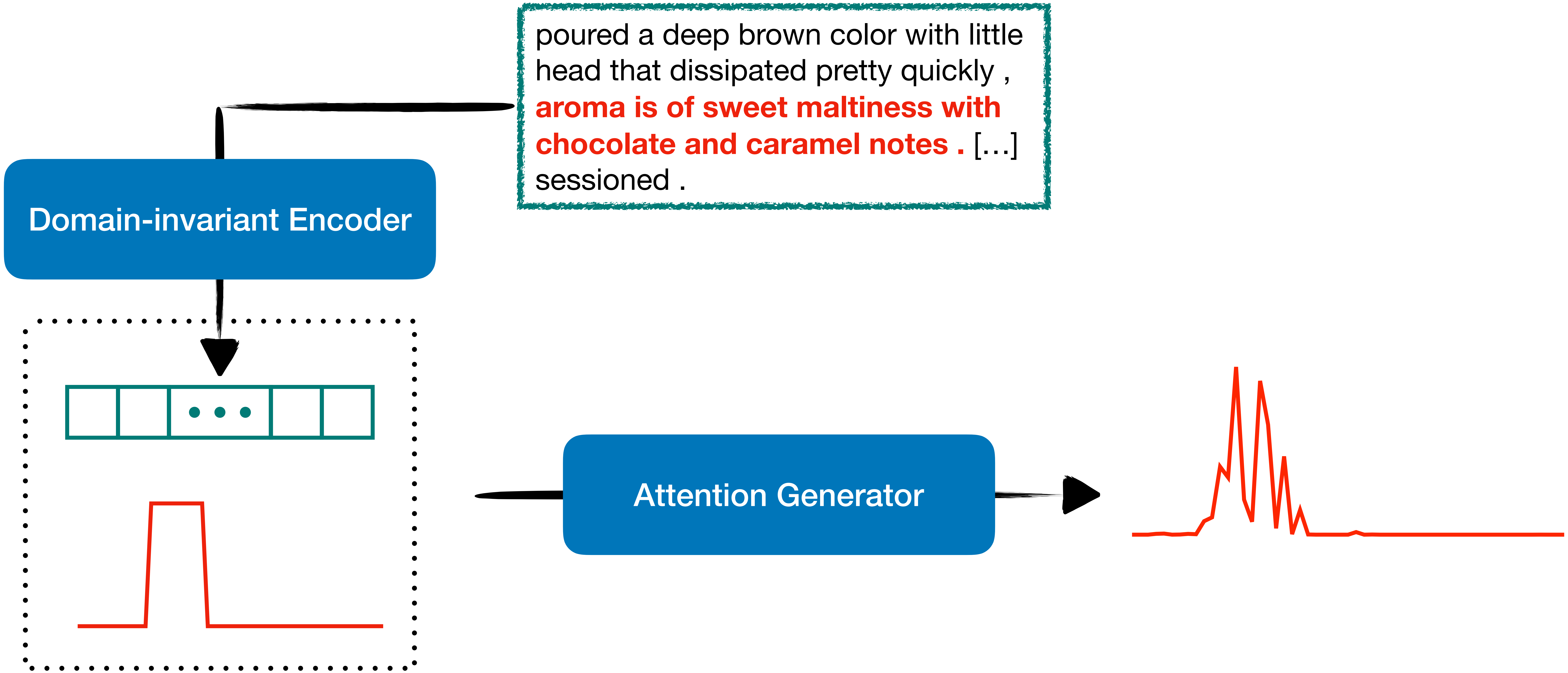
R2A Training



a nice and clean hotel to stay for business and leisure . **but the location is not good if you need public transport .** […] i never tried .

Three components are jointly optimized during training.

R2A Inference

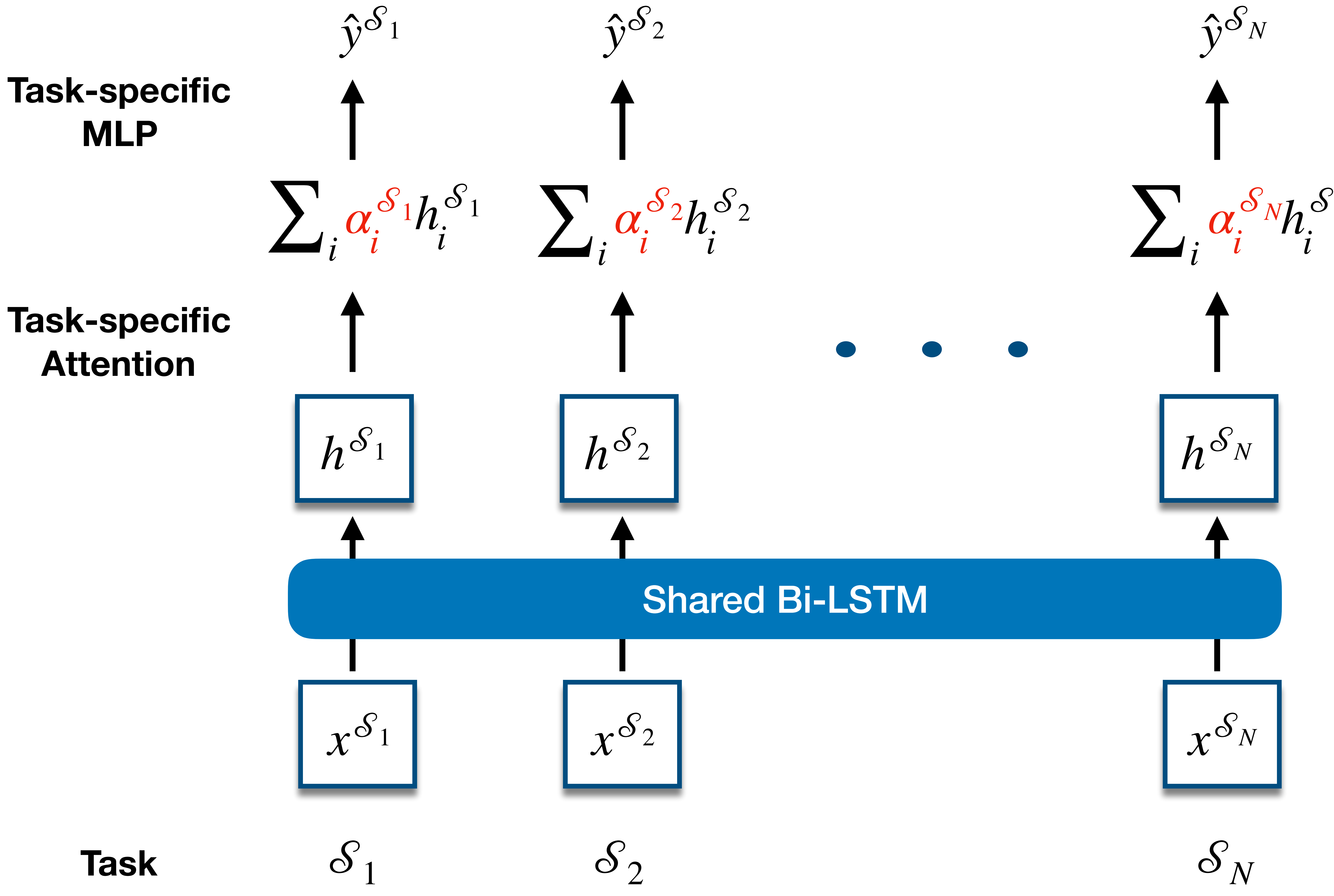


poured a deep brown color with little head that dissipated pretty quickly , **aroma is of sweet maltiness with chocolate and caramel notes .** […] sessioned .

Domain-invariant Encoder

Attention Generator

R2A: Multitask Learning



$\hat{y}^{\mathcal{S}_1}$
$\hat{y}^{\mathcal{S}_2}$
$\hat{y}^{\mathcal{S}_N}$
Task-specific MLP
$\sum_i \alpha_i^{\mathcal{S}_1} h_i^{\mathcal{S}_1}$
$\sum_i \alpha_i^{\mathcal{S}_2} h_i^{\mathcal{S}_2}$
$\sum_i \alpha_i^{\mathcal{S}_N} h_i^{\mathcal{S}_N}$
Task-specific Attention
$h^{\mathcal{S}_1}$
$h^{\mathcal{S}_2}$
$h^{\mathcal{S}_N}$
Shared Bi-LSTM
$x^{\mathcal{S}_1}$
$x^{\mathcal{S}_2}$
$x^{\mathcal{S}_N}$
Task
$\mathcal{S}_1$
$\mathcal{S}_2$
$\mathcal{S}_N$

Source tasks:

$$\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_N$$

Goal:

Generate oracle attention for each source task.

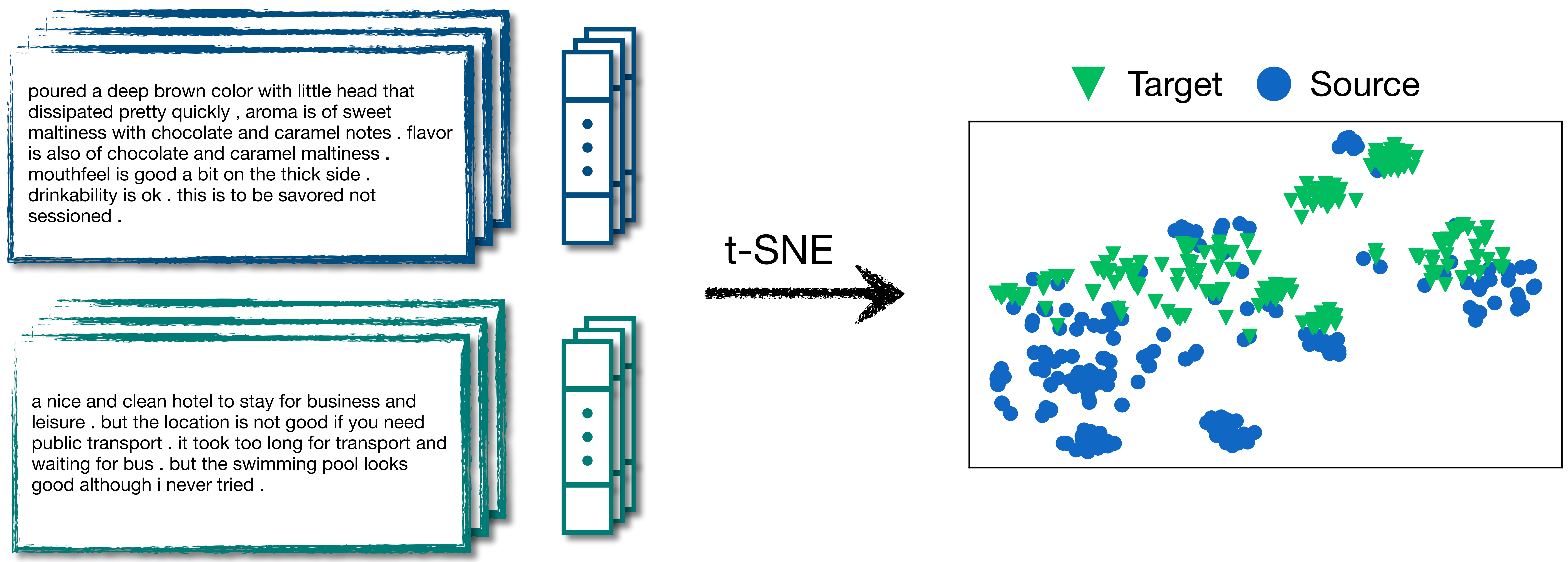
Loss:

Prediction error on all source tasks

R2A: Domain-invariant Encoder

Source Task
(beer aroma)

Target Task
(hotel cleanliness)



poured a deep brown color with little head that dissipated pretty quickly , aroma is of sweet maltiness with chocolate and caramel notes . flavor is also of chocolate and caramel maltiness . mouthfeel is good a bit on the thick side . drinkability is ok . this is to be savored not sessioned .

a nice and clean hotel to stay for business and leisure . but the location is not good if you need public transport . it took too long for transport and waiting for bus . but the swimming pool looks good although i never tried .

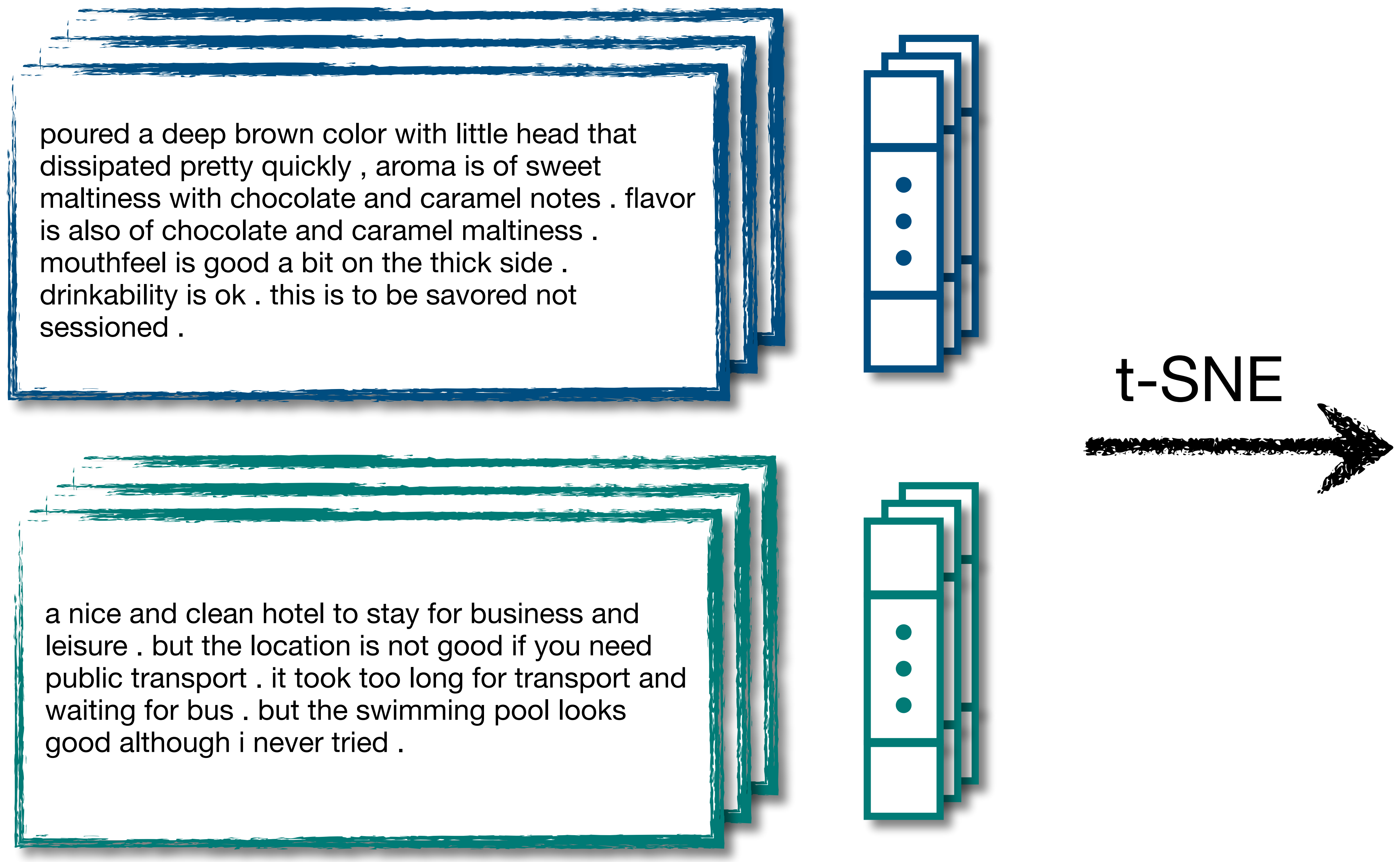
Goal:
Learn an invariant feature representation for the source and the target task.

Loss:
Wasserstein distance between source and target feature distributions.

R2A: Domain-invariant Encoder

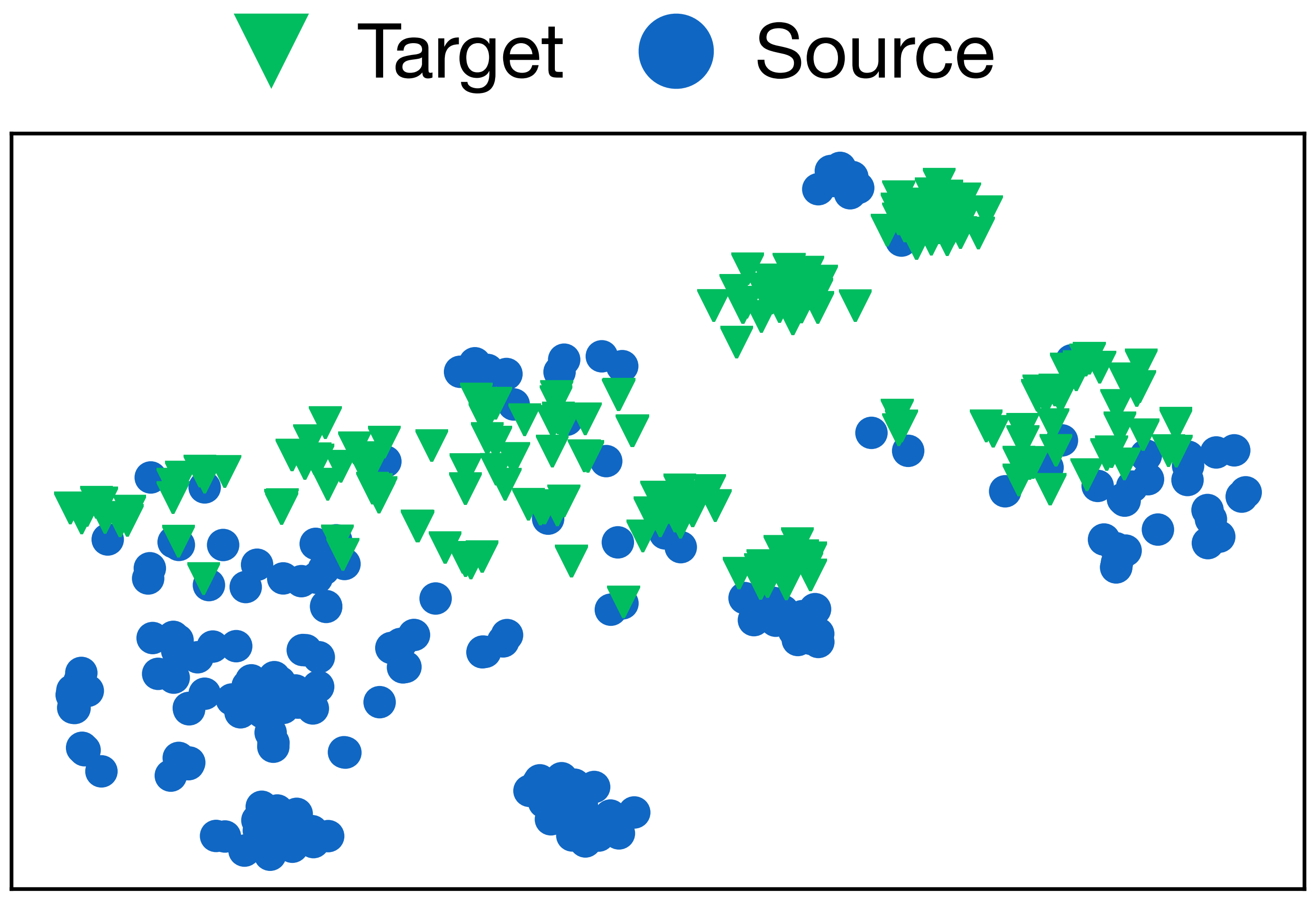
Source Task
(beer aroma)

Target Task
(hotel cleanliness)

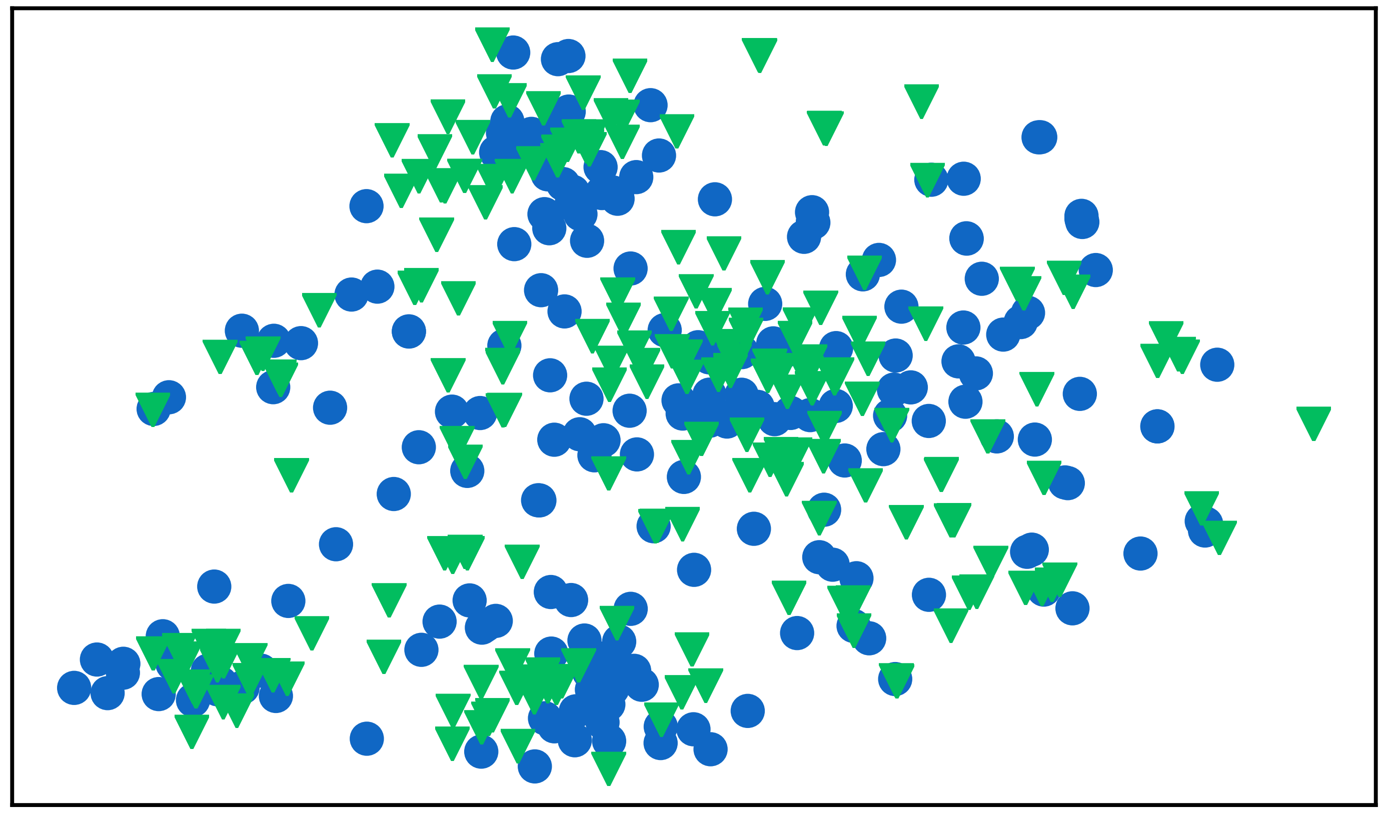


poured a deep brown color with little head that dissipated pretty quickly , aroma is of sweet maltiness with chocolate and caramel notes . flavor is also of chocolate and caramel maltiness . mouthfeel is good a bit on the thick side . drinkability is ok . this is to be savored not sessioned .

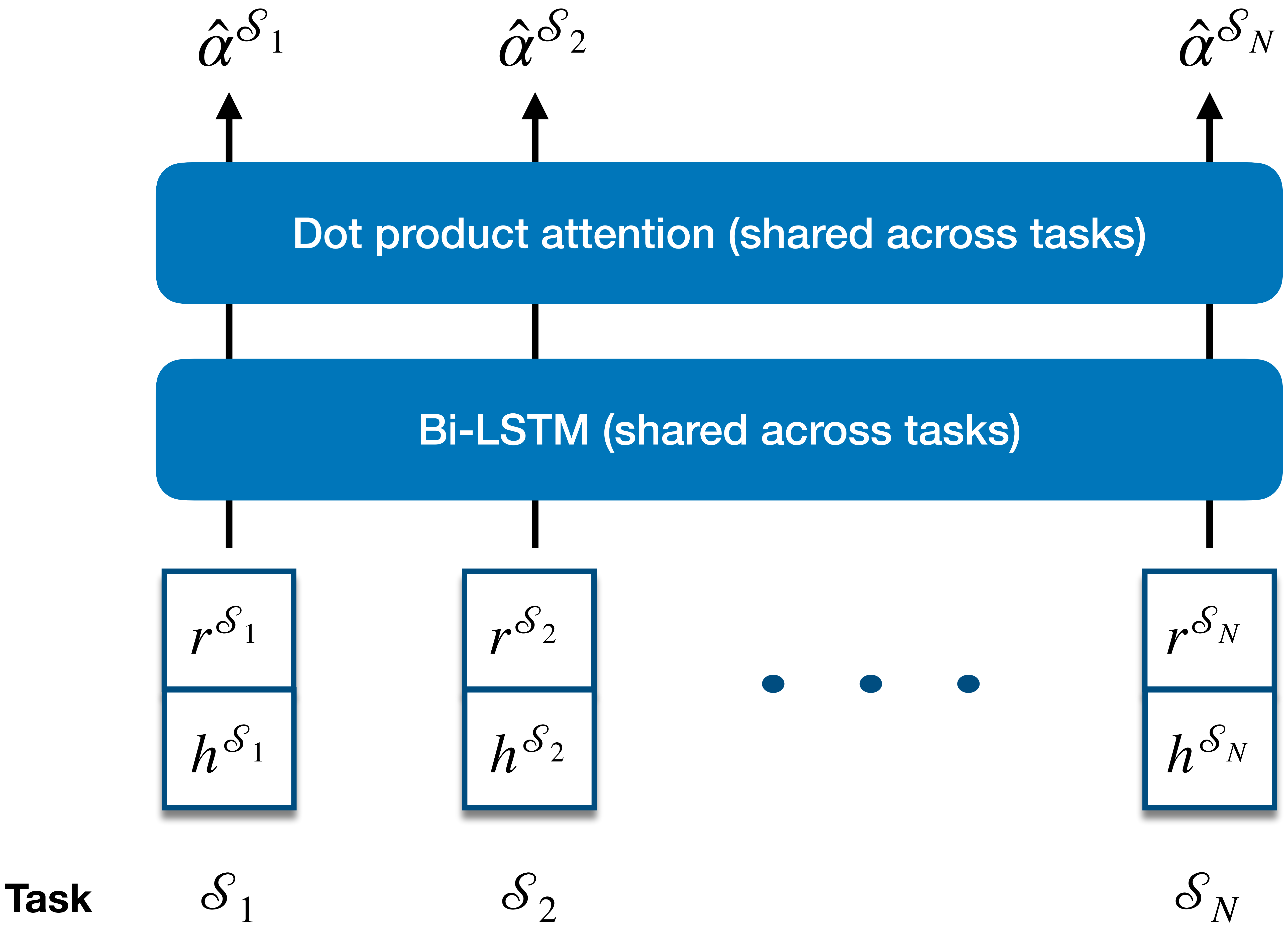
a nice and clean hotel to stay for business and leisure . but the location is not good if you need public transport . it took too long for transport and waiting for bus . but the swimming pool looks good although i never tried .



After alignment:



R2A: Attention Generator



Source tasks:

$$\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_N$$

Goal:
Predict oracle attention from rationales and the input representation.

Loss:
Distance between the generated attention $\hat{\alpha}^{\mathcal{S}_i}$ and the oracle attention $\alpha^{\mathcal{S}_i}$ (obtained from multi-task learning)

Experimental Setup

Tasks:

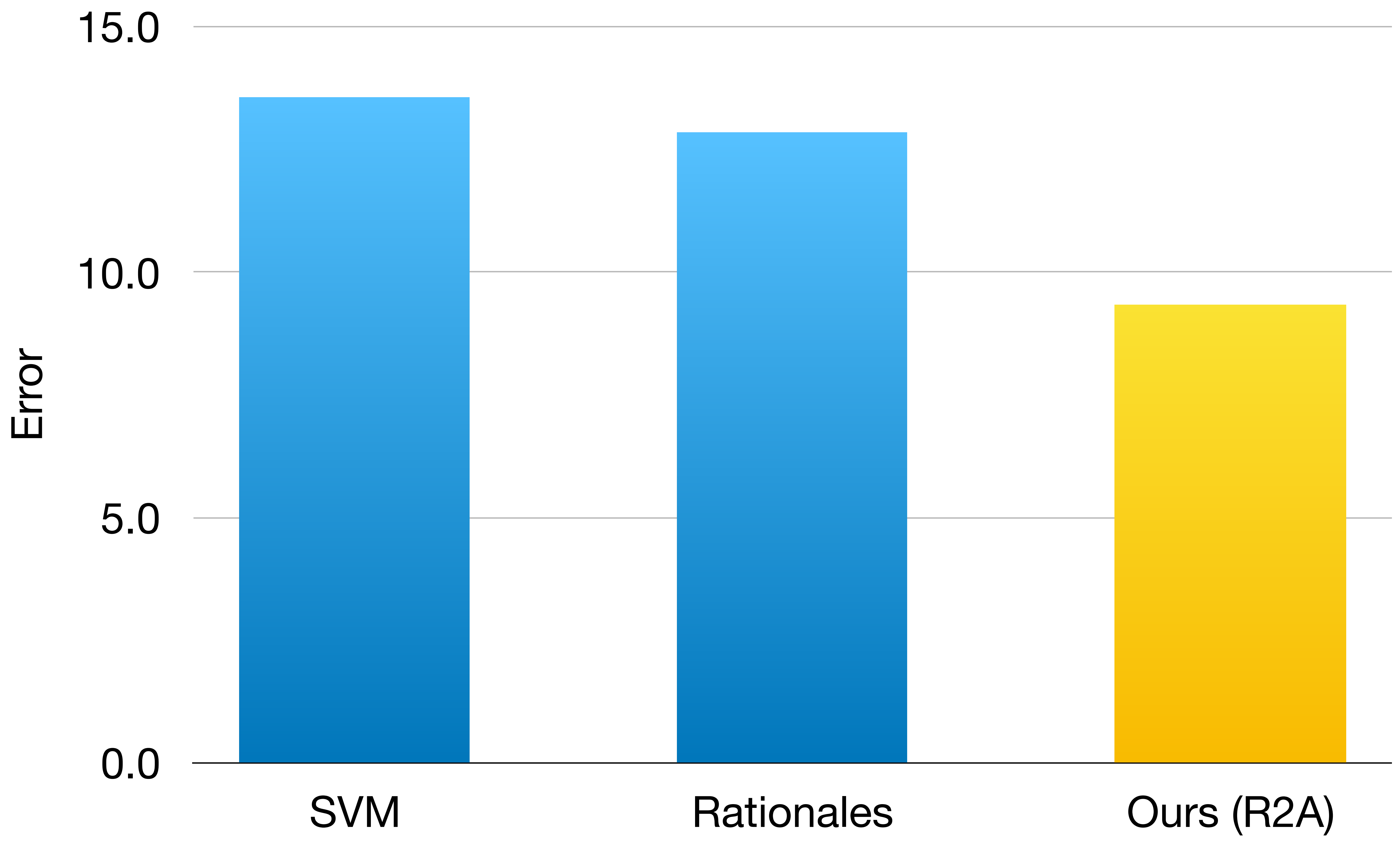
Sentiment analysis on different aspects from two domains.

Data:

BeerAdvocate review, TripAdvisor hotel review

	Tasks	Train	Test
Source	**Beer Look**	43,351	10,170
	Beer Aroma	39,825	8,772
	Beer Palate	30,041	7,152
Target	**Hotel Cleanliness**	200	12,684

Result

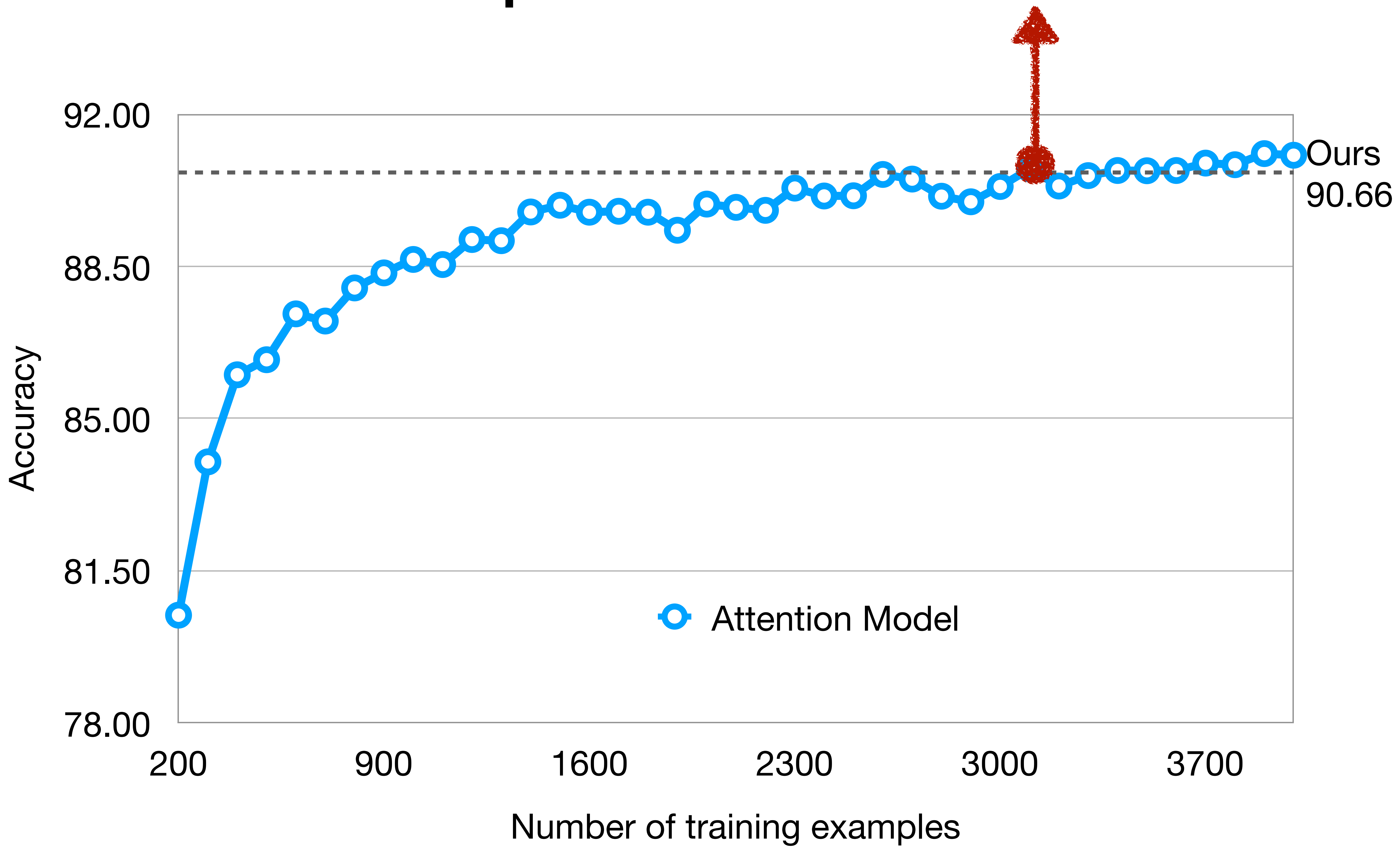


R2A as a proxy for oracle

27% error reduction!

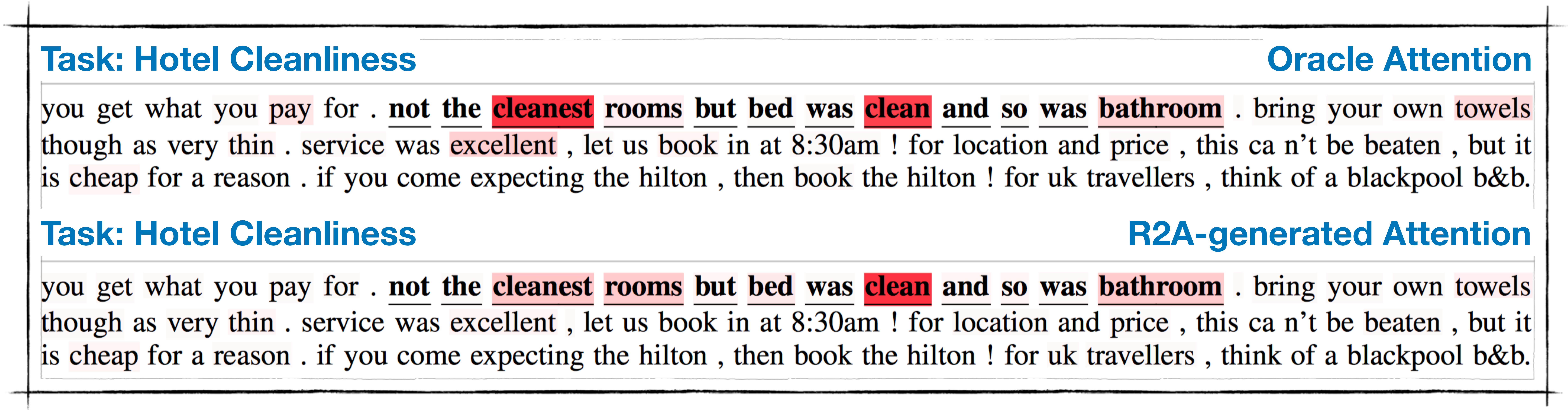
Annotating on a Budget: Rationales *vs* More Data

200 examples+rationales = 3100 examples



Same performance on 6% of the data

R2A-generated Attention *vs* Oracle Attention



Task: Hotel Cleanliness

Oracle Attention

you get what you pay for . **not the cleanest rooms but bed was clean and so was bathroom** . bring your own towels though as very thin . service was excellent , let us book in at 8:30am ! for location and price , this ca n't be beaten , but it is cheap for a reason . if you come expecting the hilton , then book the hilton ! for uk travellers , think of a blackpool b&b.

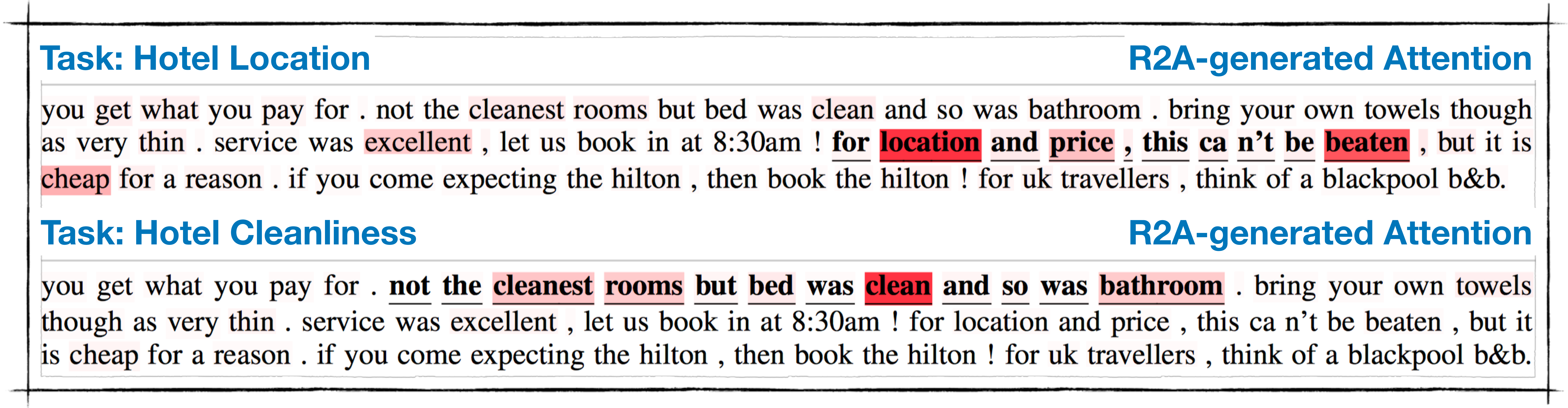
Task: Hotel Cleanliness

R2A-generated Attention

you get what you pay for . **not the cleanest rooms but bed was clean and so was bathroom** . bring your own towels though as very thin . service was excellent , let us book in at 8:30am ! for location and price , this ca n't be beaten , but it is cheap for a reason . if you come expecting the hilton , then book the hilton ! for uk travellers , think of a blackpool b&b.

R2A-generated attention mimics oracle attention

R2A-generated Attention from Different Rationales



Task: Hotel Location **R2A-generated Attention**

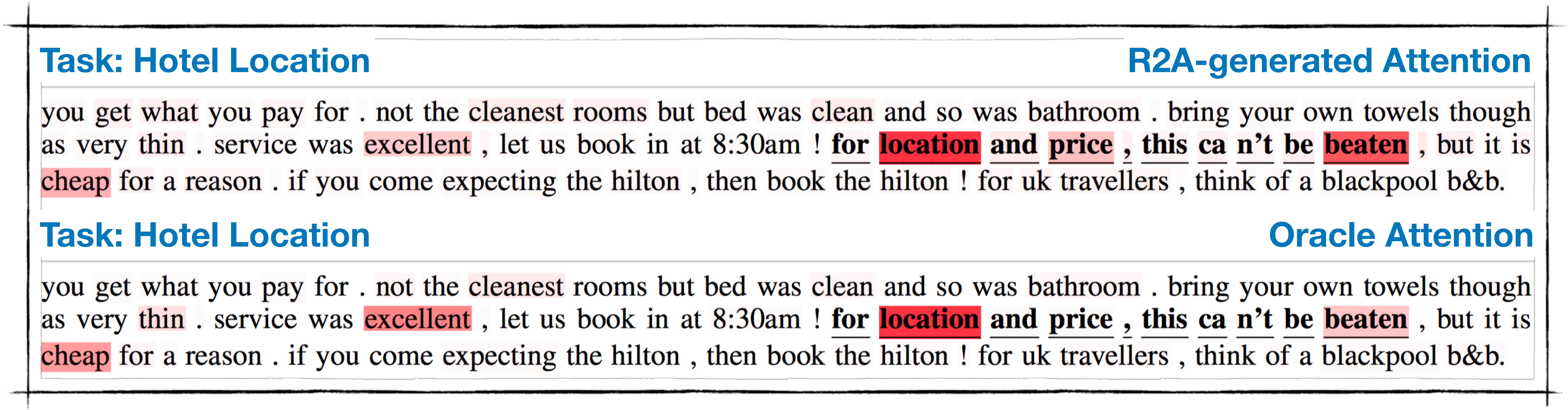
you get what you pay for . not the cleanest rooms but bed was clean and so was bathroom . bring your own towels though as very thin . service was excellent , let us book in at 8:30am ! **for location and price , this ca n't be beaten** , but it is cheap for a reason . if you come expecting the hilton , then book the hilton ! for uk travellers , think of a blackpool b&b.

Task: Hotel Cleanliness **R2A-generated Attention**

you get what you pay for . **not the cleanest rooms but bed was clean and so was bathroom** . bring your own towels though as very thin . service was excellent , let us book in at 8:30am ! for location and price , this ca n't be beaten , but it is cheap for a reason . if you come expecting the hilton , then book the hilton ! for uk travellers , think of a blackpool b&b.

R2A-generated attention changes according to the input rationales.

R2A-generated Attention *vs* Oracle Attention



Task: Hotel Location

R2A-generated Attention

you get what you pay for . not the cleanest rooms but bed was clean and so was bathroom . bring your own towels though as very thin . service was excellent , let us book in at 8:30am ! **for location and price , this ca n't be beaten** , but it is cheap for a reason . if you come expecting the hilton , then book the hilton ! for uk travellers , think of a blackpool b&b.

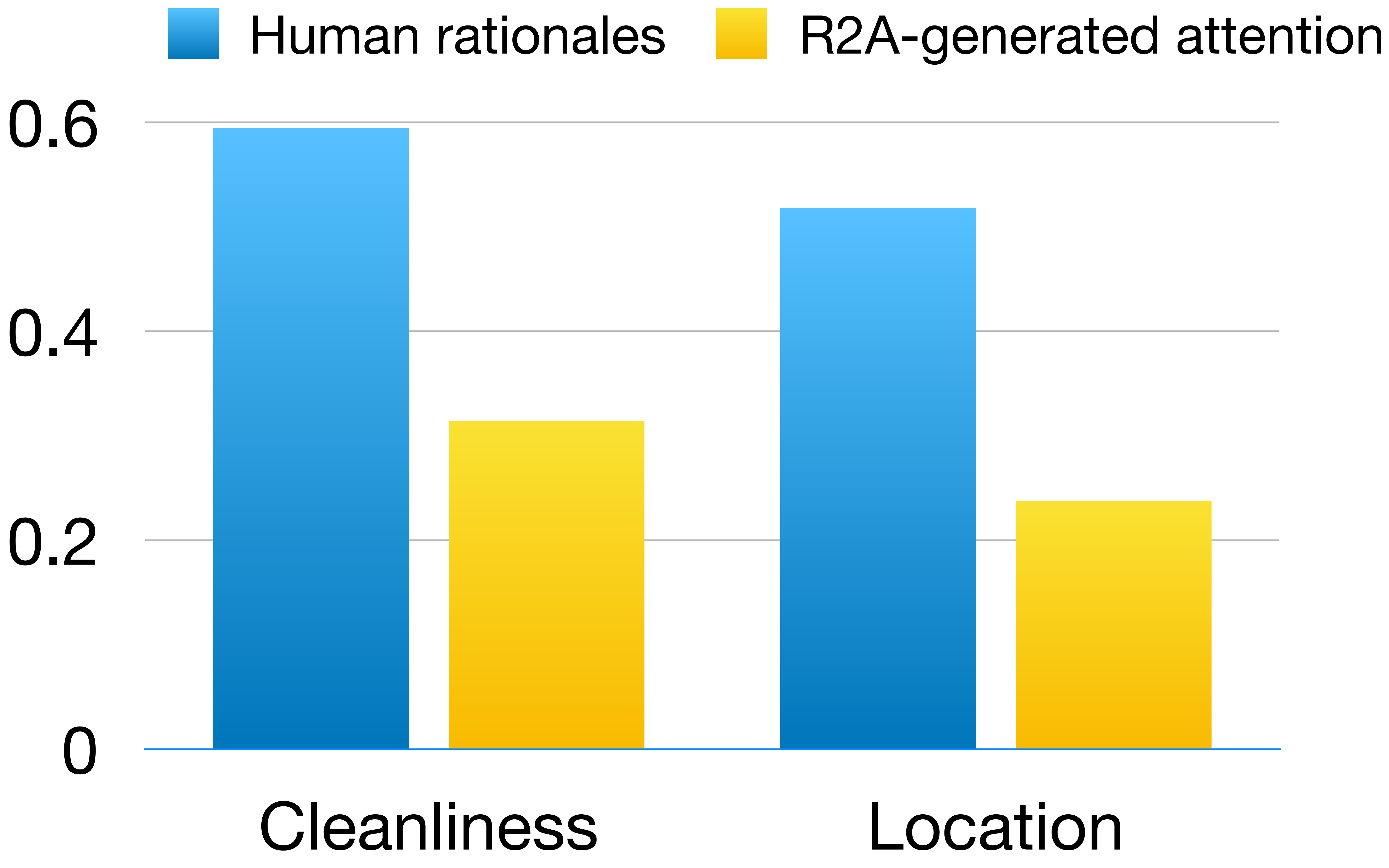
Task: Hotel Location

Oracle Attention

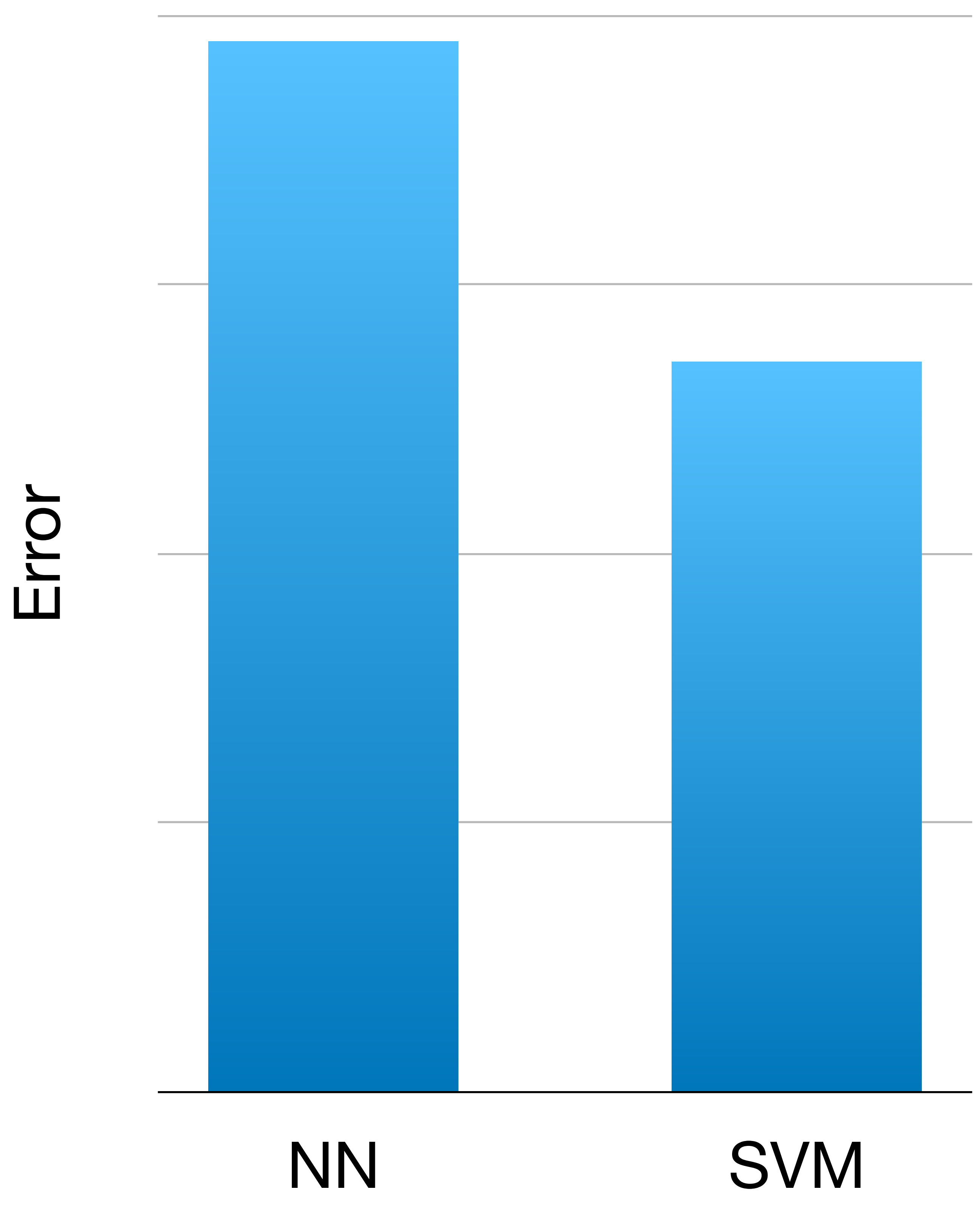
you get what you pay for . not the cleanest rooms but bed was clean and so was bathroom . bring your own towels though as very thin . service was excellent , let us book in at 8:30am ! **for location and price , this ca n't be beaten** , but it is cheap for a reason . if you come expecting the hilton , then book the hilton ! for uk travellers , think of a blackpool b&b.

R2A-generated attention mimics oracle attention

Cosine Distance to Oracle Attention



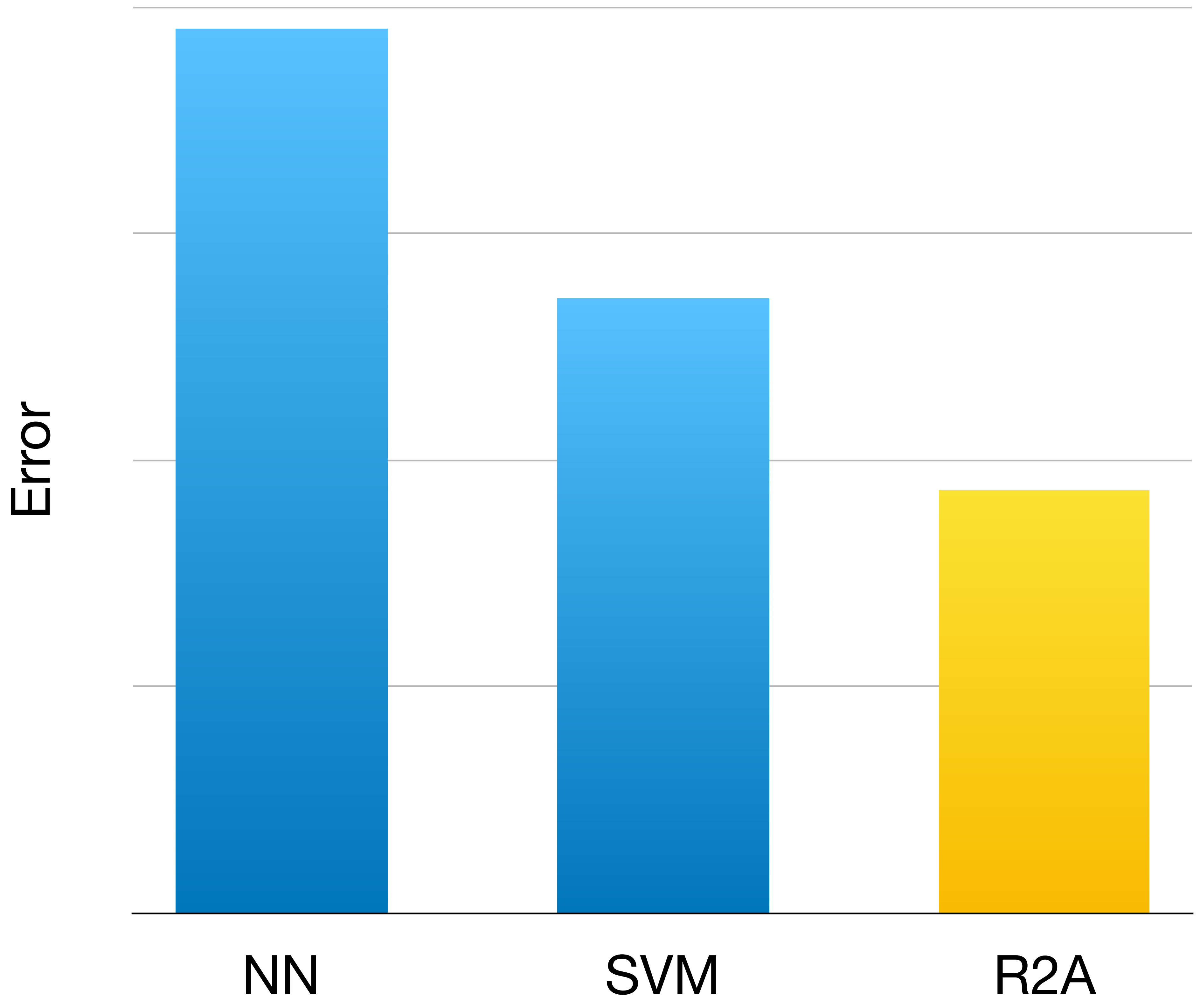
R2A-generated attention is closer to the oracle.



Training data: **200** instances

Can NN do better on small training sets?

Conclusions



Training data: **200** instances

Yes, it can.

Code & data: https://github.com/YujiaBao/R2A

Thank you