Deriving Machine Attention from Human Rationales

Yujia Bao¹

Shiyu Chang², Mo Yu², Regina Barzilay¹

¹Computer Science and Artificial Intelligence Lab, MIT
²MIT-IBM Watson AI Lab, IBM Research
Neural Networks in Low-resource Scenario

Can NN do better on small training sets?

Training data: 200 instances
Human Rationales can Help


- **Rationales are useful for training SVMs**
- **Limited benefits for neural models**

Conventional Supervised Training

Supervised Training with Rationales

Input

Label

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 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.

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

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

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.
Observations:

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

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.


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.

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

**Source tasks:** rationales are generated automatically

---

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.
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.
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.**
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.
poured a deep brown color with little head that dissipated pretty quickly, aroma is of sweet maltiness with chocolate and caramel notes. [...] sessioned.
R2A: Multitask Learning

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

\[
\sum_i \alpha_i^{\mathcal{S}_1} h_i^{\mathcal{S}_1} \quad \sum_i \alpha_i^{\mathcal{S}_2} h_i^{\mathcal{S}_2} \quad \sum_i \alpha_i^{\mathcal{S}_N} h_i^{\mathcal{S}_N}
\]

Task-specific MLP

Task-specific Attention

Shared Bi-LSTM

Task
\( \mathcal{S}_1 \) \( \mathcal{S}_2 \) \( \mathcal{S}_N \)
R2A: Domain-invariant Encoder

Source Task
(beer aroma)

Target Task
(hotel cleanliness)

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:

t-SNE
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

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beer Look</td>
<td>43,351</td>
<td>10,170</td>
</tr>
<tr>
<td>Beer Aroma</td>
<td>39,825</td>
<td>8,772</td>
</tr>
<tr>
<td>Beer Palate</td>
<td>30,041</td>
<td>7,152</td>
</tr>
<tr>
<td>Hotel Cleanliness</td>
<td>200</td>
<td>12,684</td>
</tr>
</tbody>
</table>
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 \( \text{vs} \) 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 can’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 can’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

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 can’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

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 can’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 can’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.

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 can’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.
Can NN do better on small training sets?

Training data: 200 instances
Conclusions

Training data: 200 instances

Yes, it can.

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