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
{yujia, regina}@csail.mit.edu, shiyu.chang@ibm.com, yum@us.ibm.com

Abstract

Attention-based models are successful when trained on large amounts of data. In this paper, we demonstrate that even in the low-resource scenario, attention can be learned effectively. To this end, we start with discrete human-annotated rationales and map them into continuous attention. Our central hypothesis is that this mapping is general across domains, and thus can be transferred from resource-rich domains to low-resource ones. Our model jointly learns a domain-invariant representation and induces the desired mapping between rationales and attention. Our empirical results validate this hypothesis and show that our approach delivers significant gains over state-of-the-art baselines, yielding over 15% average error reduction on benchmark datasets.[1]

1 Introduction

Attention-based models have become architectures of choice for many NLP tasks. In addition to significant performance gains, these models are attractive, as attention is often used as a proxy for human interpretable rationales. Their success, however, is conditioned on access to large amounts of training data. To make these models applicable in low-resource scenarios, we utilize this connection in the opposite direction. Specifically, we propose an approach to map human rationales to high-performing attention, and use this attention to guide models trained in low-resource scenarios.

The notions of rationale and attention are closely related. Both of them highlight word importance for the final prediction. In the case of rationale, the importance is expressed as a hard selection, while attention provides a soft distribution over the words. Figure 1 illustrates this relatedness. One obvious approach to improve low-resource performance is to directly use human rationales as a supervision for attention generation. The implicit assumption behind this method is that machine-generated attention should mimic human rationales. However, rationales on their own are not adequate substitutes for machine attention. Instead of providing a soft distribution, human rationales only provide the binary indication about relevance. Furthermore, rationales are subjectively defined and often vary across annotators. Finally, human rationales are not customized for a given model architecture. In contrast, machine attention is always derived as a part of a specific model architecture.

Task: Hotel location — label: negative

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 .

Task: Beer aroma — label: positive

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 .

Figure 1: Examples of rationales versus oracle attention. Words are highlighted according to their relative attention scores. Human rationales are shown in bold with underlines.

To further understand this connection, we empirically compare models informed by human rationales and those by high-quality attention. To obtain the latter, we derive an "oracle" attention using a large amount of annotations. This "oracle" attention is then used to guide a model that only has access to a small subset of this training data. Not only does this model outperform the oracle-free variant, but it also yields substantial gains over its counterpart trained with human ra-

[1]Our code and data are available at https://github.com/YujiaBao/R2A.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1903–1913
Brussels, Belgium, October 31 - November 4, 2018.

tionales — 89.98 % vs 85.22 % average accuracy on three aspects of hotel review (see Section 4 for details). In practice, however, this "oracle" attention is not available. To employ this method, we need to find a way to obtain a substitute for the "oracle" attention.

In this paper, we show how to achieve this goal using rationales. Specifically, we learn a mapping from human rationales to high-quality attention (R2A). We hypothesize that this mapping is generalizable across tasks and thus can be transferred from resource-rich tasks.[2] Figure 1 illustrates that in both tasks, attention weighs rationale words in a similar fashion: highlighting task-specific nouns and adjectives, while downplaying functional words. To learn and apply this mapping we need access to rationales in both source and target tasks. In the target task, we assume rationales are provided by humans. In the source task(s), collecting rationales at scale is infeasible. Therefore, we use machine-generated rationales (Lei et al., 2016) as a proxy.

Our R2A model consists of three components. The first one is an attention-based model for the source task(s) that provides supervision for attention generation. The second component focuses on learning a domain-invariant representation to support transfer. The third component combines this invariant representation and rationales together to generate the attention. These three components are trained jointly to optimize the overall objective. Once the model is trained, we apply it to the target task to generate attention from human rationales. This attention is consequently used to supervise the training of the target classifier.

We evaluate our approach on two transfer settings: aspect transfer within single domain and domain transfer across multiple domains. Our experiments demonstrate that our approach delivers significant performance improvements over the baselines. For instance, the average error reduction over the best baseline in domain transfer is over 15%. In addition, both qualitative and quantitative analyses confirm that our R2A model is capable of generating high-quality attention for target tasks.

[2]In this paper, we consider a more general setting where one domain contains multiple tasks. Also, we assume having one source domain. However, our proposed method is a general framework and can be easily adapted to problems with multiple source domains.

2 Related Work

Attention-based models Attention has been shown to be effective when the model is trained on large amounts of training data (Bahdanau et al., 2014; Luong et al., 2015; Rush et al., 2015; Yang et al., 2016; Lin et al., 2017; Chen et al., 2017; Vaswani et al., 2017). In this setting, typically no additional supervision is required for learning the attention. Nevertheless, further refining attention by extra supervision has been shown to be beneficial. Examples include using word alignments to learn attention in neural machine translation (Liu et al., 2016), employing argument words to supervise attention in event detection (Liu et al., 2017), utilizing linguistically-motivated annotations to guide attention in constituency parsing (Kamigaito et al., 2017). These supervision mechanisms are tailored to specific applications. In contrast, our approach is based on the connection between rationales and attention, and can be used for multiple applications.

Rationale-based models Zaidan et al. (2007) was the first to explore the value of rationales in low-resource scenarios. They hypothesize that the model confidence should decrease when the rationale words are removed from the inputs, and validate this idea for linear models. Recent work (Zhang et al., 2016) explores the potential of integrating rationales with more complex neural classifiers. In their model, human rationales are directly used to guide the sentence-level attention for a CNN-based classifier. To reach good performance, their model still requires a sufficient amount of training data. Our work differs from theirs as we discern the intrinsic difference between human rationales and machine attention. Moreover, we learn a model to map human rationales into high-quality attention so as to provide a richer supervision for low-resource models.

Transfer learning When labeled data on the target task is available, existing approaches typically transfer the knowledge by either fine-tuning an encoder trained on the source tasks(s) (Conneau et al., 2017; Peters et al., 2018) or multi-task learning on all tasks with a shared encoder (Collobert et al., 2011). In this paper, we explore the transferability of the task-specific attention through human rationales. We believe this will further assist learning in low-resource scenarios.

Our work is also related to unsupervised domain

adaptation, as the R2A model has never seen any target annotations during training. Existing methods commonly adapt the classifier by aligning the representations between the source and target domains (Glorot et al., 2011; Chen et al., 2012; Zhou et al., 2016; Ganin et al., 2016; Zhang et al., 2017). In contrast, our model adapts the mapping from rationales to attention; thus after training, it can be applied to different target tasks.

3 Method

Problem formulation We assume that we have N source tasks $\{\mathcal{S}_i\}_{i=1}^{N}$, where each of them has sufficient amounts of labeled examples. Using existing methods (Lei et al., 2016), we can generate rationales for each source example automatically (see Appendix 1 for details). In the target task $\mathcal{T}$, we only have a limited amount of labeled examples with large amounts of unlabeled data. For those labeled examples, we assume access to human-annotated rationales.

Overview Our goal is to improve classification performance on the target task by learning a mapping from human rationales to high-quality machine attention (R2A). Given the scarcity of our target data, we learn this mapping on resource rich tasks where high-quality attention can be readily obtained during training. Next, the mapping between rationales and attention derived from the source tasks is exported into the target task. To enable this transfer, models have to operate over an invariant representation which we construct via an adversarial objective. Once the mapping is derived, we can translate human rationales in the target task into high-quality attention. This generated attention is then used to provide additional training signal for an attention-based classifier for the target task. The overall pipeline is shown in Figure 2.

Alternatively, we can view the R2A mapping as a meta model that produces a prior over the attention distribution across different tasks.

Model architecture Figure 3 illustrates the architecture of our R2A model, which consists of three components.

- **Multi-task learning** In order to learn the R2A mapping, we need annotation for the attention. This module generates high-quality attention as an intermediate result by minimizing the prediction error on the source tasks (Section 3.1).

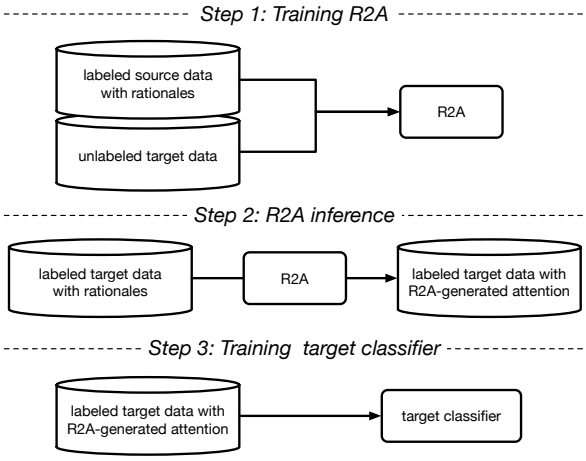


Figure 2: Overall pipeline of our approach (Section 3.4). The R2A mapping is learned from labeled source data and unlabeled target data. Then we applied it to the target task to derive attention based on human rationales. Finally, a target classifier is trained under the supervision of both the annotated labels and the R2A-generated attention.

- **Domain-invariant encoder** This module aims to transform the contextualized representation obtained from the first module into a domain-invariant version. We achieve this goal through domain adversarial training over the source data and the unlabeled target data (Section 3.2).

- **Attention generation** This module learns to predict the intermediate attention obtained from the first module based on the domain-invariant representation and the rationales (Section 3.3).

3.1 Multi-task learning

The goal of the multi-task learning module is to learn good attention for each source task. This learned attention will be used later to supervise the attention generation module. This module takes the input text from the source tasks and predicts the labels. To accomplish the previously stated goal, we minimize the prediction error over all labeled source data.

Let (x^t, y^t) be a training instance from any source task $t \in \{\mathcal{S}_1, \ldots \mathcal{S}_N\}$. We first encode the input sequence x^t into hidden states: $h^t = \mathbf{enc}(x^t)$, where $\mathbf{enc}$ is a bi-directional LSTM (Hochreiter and Schmidhuber, 1997) that is shared across all source tasks. For each position i, the dense vector h_i^t encodes the content and context information of the word x_i^t. We then pass the

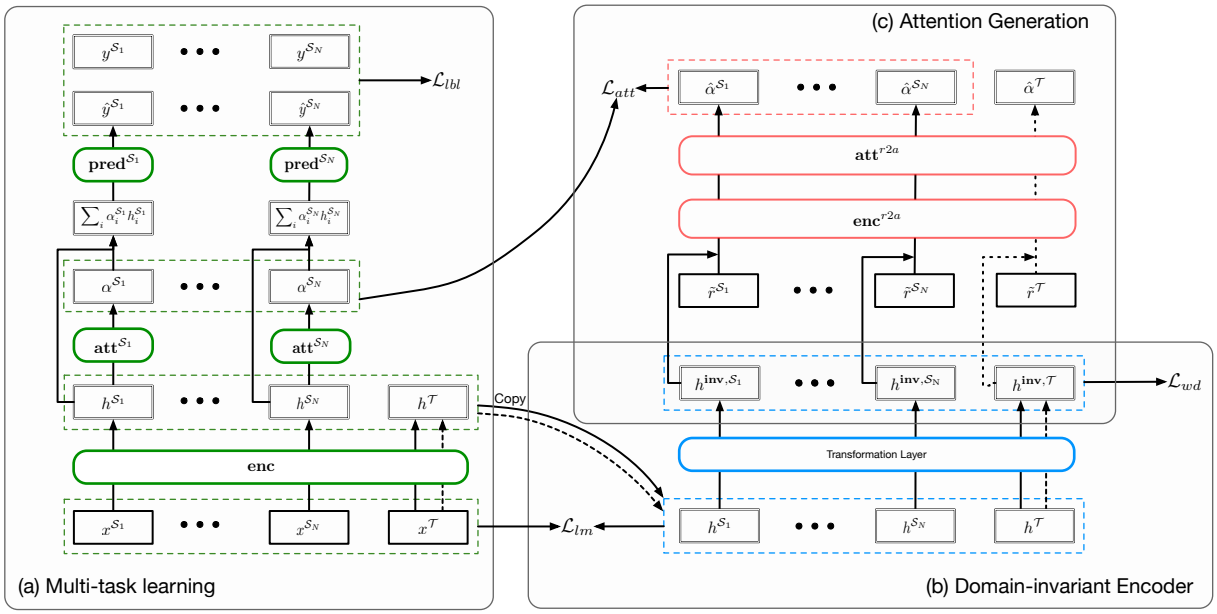


Figure 3: Architecture of the R2A model. The model is comprised of (a) a multi-task learning component, (b) a domain-invariant encoder, and (c) an attention generation component. Solid arrows denote computations for training, while dotted arrows denote computations for inference.

sequence h^t on to a task-specific attention module to produce attention $\alpha^t = \mathbf{att}^t(h^t)$ as follows:

$$\tilde{h}_i^t = \tanh(W_{\mathbf{att}}^t h_i^t + b_{\mathbf{att}}^t),$$
$$\alpha_i^t = \frac{\exp(\langle \tilde{h}_i^t, q_{\mathbf{att}}^t \rangle)}{\sum_j \exp(\langle \tilde{h}_j^t, q_{\mathbf{att}}^t \rangle)},$$

where $\langle \cdot, \cdot \rangle$ denotes inner product and $W_{\mathbf{att}}^t$, $b_{\mathbf{att}}^t$, $q_{\mathbf{att}}^t$ are learnable parameters. We predict the label of x^t using the weighted sum of its contextualized representation: $\hat{y}^t = \mathbf{pred}^t(\sum_i \alpha_i^t h_i^t)$, where $\mathbf{pred}^t$ is a task-specific multi-layer perceptron. We train this module to minimize the loss, $\mathcal{L}_{lbl}$, between the prediction and the annotated label for all source tasks. We use cross entropy loss for classification tasks and mean square loss for regression tasks.

3.2 Domain-invariant encoder

Supplied with large amounts of source data and unlabeled target data, this module has two goals: 1) learning a general encoder for both source and target corpora, and 2) learning domain-invariant representation. This module enables effective transfer—especially in the presence of significant variance between the source and target domains. We achieve the first goal by optimizing a language modeling objective and the second goal by minimizing the Wasserstein distance between the source and target distribution.

Let x be an input sequence, and $h \triangleq [\overrightarrow{h}; \overleftarrow{h}]$ be its corresponding contextualized representation obtained from **enc**. Here, $\overrightarrow{h}$ and $\overleftarrow{h}$ denote the output sequence of the forward and backward LSTM, respectively. In order to support transfer, **enc** should be general enough to effectively represent both source and target corpora. For this reason, we ground the encoder by a language modeling component (Bengio et al., 2003; Mikolov et al., 2011). Specifically, we employ two Softmax classifiers to predict the word x_i based on $\overrightarrow{h}_{i-1}$ and $\overleftarrow{h}_{i+1}$ respectively. We minimize the cross-entropy loss $\mathcal{L}_{lm}$ over all source data and unlabeled target data.

The representation h is domain-specific as it is trained to encode useful features for language modeling and the source tasks. To obtain an invariant representation, we employ a transformation layer and propose to align the transformed representation so that it is not distinguishable whether it comes from the source or the target. Specifically, we transform the representation h_i at each position i linearly and obtain

$$h_i^{\mathbf{inv}} = W_{\mathbf{inv}} h_i + b_{\mathbf{inv}},$$

where $W_{\mathbf{inv}}$ and $b_{\mathbf{inv}}$ are learnable parameters. We minimize the Wasserstein distance (Arjovsky et al., 2017) between the distribution of $h^{\mathbf{inv}}$ from the source and the one from the target, denoted as $\mathbb{P}_{\mathcal{S}}$ and $\mathbb{P}_{\mathcal{T}}$, respectively. Since $h^{\mathbf{inv}}$ is a sequence

of variable length, L, we summarize it by its first and last element via concatenation, $[h_1^{\mathbf{inv}}; h_L^{\mathbf{inv}}]$. The training objective is defined as:

$$\mathcal{L}_{wd} = \sup_{\|f\|_L \leq K} \mathbb{E}_{h^{\mathbf{inv}} \sim \mathbb{P}_{\mathcal{S}}} \left[f([h_1^{\mathbf{inv}}; h_L^{\mathbf{inv}}]) \right] - \mathbb{E}_{h^{\mathbf{inv}} \sim \mathbb{P}_{\mathcal{T}}} \left[f([h_1^{\mathbf{inv}}; h_L^{\mathbf{inv}}]) \right],$$

where the supremum is over all K-Lipschitz scalar functions f. Following Gulrajani et al. (2017), we approximate f by a multi-layer perceptron, and use gradient penalty to fulfill the Lipschitz constraint.

3.3 Attention generation

The goal of this module is to generate high-quality attention for each task. This module combines the domain-invariant representation together with task-specific rationales as its input and predicts task-specific attention scores. We minimize the distance between the predicted attention and the intermediate attention obtained from the multi-task learning module.

For any source task $t \in \{\mathcal{S}_i\}_{i=1}^N$, we denote r^t as the task-specific rationales corresponding to the input text x^t, and denote $h^{\mathbf{inv},t}$ as the domain-invariant representation of x^t. For each position i, we first concatenate r_i^t with the frequency of x_i^t occurring as a rationale from all training examples of this task. We denote this augmented sequence as $\tilde{r}^t$. This frequency term provides the unigram likelihood of each word being a rationale for the task. Then we employ a sequence encoder $\mathbf{enc}^{r2a}$ and an attention module $\mathbf{att}^{r2a}$ to predict the attention scores:

$$u^t = \mathbf{enc}^{r2a}([h^{\mathbf{inv},t}; \tilde{r}^t]),$$
$$\tilde{u}_i^t = \tanh(W_{\mathbf{att}}^{r2a} u_i^t + b_{\mathbf{att}}^{r2a}),$$
$$\hat{\alpha}_i^t = \frac{\exp(\langle \tilde{u}_i^t, q_{\mathbf{att}}^{r2a} \rangle)}{\sum_j \exp(\langle \tilde{u}_j^t, q_{\mathbf{att}}^{r2a} \rangle)},$$

where $W_{\mathbf{att}}^{r2a}$, $b_{\mathbf{att}}^{r2a}$ and $q_{\mathbf{att}}^{r2a}$ are learnable parameters, and both $\mathbf{enc}^{r2a}$ and $\mathbf{att}^{r2a}$ are shareable across all tasks. We minimize the distance between $\hat{\alpha}^t$ and the α^t obtained from the first multi-task learning module over all source data:

$$\mathcal{L}_{att} = \sum_{(\alpha^t, \hat{\alpha}^t), t \in \{\mathcal{S}_i\}_{i=1}^N} \mathrm{d}(\alpha^t, \hat{\alpha}^t),$$

where $\mathrm{d}(\cdot, \cdot)$ can be any distance metric. In this paper, we consider a soft-margin cosine distance:

$$\mathrm{d}(a, b) \triangleq \max(0, 1 - \cos(a, b) - 0.1),$$

where $\cos(\cdot, \cdot)$ denotes the cosine similarity.

3.4 Pipeline

Training R2A We train the three aforementioned modules jointly using both the source data and the unlabeled target data. The overall objective is given by:

$$\mathcal{L} = \mathcal{L}_{lbl} + \lambda_{att}\mathcal{L}_{att} + \lambda_{lm}\mathcal{L}_{lm} + \lambda_{wd}\mathcal{L}_{wd}. \quad (1)$$

The λs are hyper-parameters that control the importance of each training objective and $\mathcal{L}$s represent the corresponding loss functions.

R2A inference Once the R2A model is trained, we can generate attention for each labeled target example based on its human-annotated rationales.

Training target classifier When testing the performance on the target task, of course, we are neither provided with labels nor rationales. In order to make predictions for the target task, we train a new classifier under the supervision of both the labels and the R2A-generated attention. Specifically, this target classifier shares the same architecture as the source one in the multi-task learning module. We minimize the prediction loss, $\mathcal{L}_{lbl}^{\mathcal{T}}$, on the labeled target data together with the cosine distance, $\mathcal{L}_{att}^{\mathcal{T}}$, between the R2A-generated attention and the attention generated by this target classifier. The joint objective for this target classifier is defined as

$$\mathcal{L} = \mathcal{L}_{lbl}^{\mathcal{T}} + \lambda_{att}^{\mathcal{T}}\mathcal{L}_{att}^{\mathcal{T}}, \quad (2)$$

where $\lambda_{att}^{\mathcal{T}}$ controls the importance of $\mathcal{L}_{att}^{\mathcal{T}}$. For better transfer, we initialize the encoder in the target classifier as $\mathbf{enc}$ from the trained R2A model.

4 Experimental Setup

4.1 Datasets

We evaluate our approach on two transfer settings: transfer among aspects within the same domain and transfer among different domains.

Aspect transfer We first consider the transfer problem between multiple aspects of one domain: beer review. We use a subset of the BeerAdvocate[3] review dataset (McAuley et al., 2012) introduced by Lei et al. (2016). This dataset contains reviews with ratings (in the scale of $[0, 1]$) from three aspects of the beer: look, aroma and palate. We treat

[3] https://www.beeradvocate.com

Beer Aspects	Source Train	Source Dev	Target Train‡	Target Dev	Target Test
Look	43,351	10,170	200	200	4,014
Aroma	39,825	8,772	200	200	4,212
Palate	30,041	7,152	200	200	3,804

Table 1: Statistics of the beer review dataset. ‡ denotes data with human-annotated rationales.

Hotel Aspects	Target Train‡	Target Dev	Target Test	Target Unlabeled
Location	200	200	1,808	14,472
Cleanliness	200	200	12,684	14,472
Service	200	200	18,762	14,472

Table 2: Statistics of the hotel review dataset. ‡ denotes data with human-annotated rationales.

any two aspects as the source and the other one as the target. We consider a classification setting for each target task. Specifically, reviews with ratings ≤ 0.4 are labeled as negative and those with ≥ 0.6 are labeled as positive. We form our dataset by randomly selecting an equal number of positive and negative examples. Then we randomly select 200 examples and ask human annotators to provide rationales (see Appendix 2 for details). These 200 examples are treated as our labeled training data for the target aspect. Unlabeled target data is not required since both source and target tasks are from the same domain. Table 1 summarizes the statistics of the beer review dataset.

Domain transfer Our second experiment focuses on domain transfer from beer reviews to different aspects of hotel reviews. We use the TripAdvisor[4] hotel review dataset (Wang et al., 2010), with the following three aspects as our transfer target: location, cleanliness, and service. For each aspect, reviews with ratings > 3 are labeled as positive and those with < 3 are labeled as negative. Similarly, we collect human rationales for 200 examples and treat them as our training data (see Appendix 2 for details). Table 2 summarizes the statistics of the hotel review dataset. In this experiment, data from all three aspects of the beer reviews are treated as the source tasks.

4.2 Baselines

We compare our approach (OURS) with four types of baselines:

[4] https://www.tripadvisor.com

Basic classifier We train a linear SVM using bag-of-ngrams representation on the labeled target data. We combine uni-gram, bi-grams, and tri-grams as features and use tf-idf weighting.

Rationale augmented classifiers We evaluate two previous methods that incorporate human rationales during training: 1) rationale augmented SVM (RA-SVM) (Zaidan et al., 2007), an SVM-based model that utilizes human rationales to regularize the decision boundary of the classifier; 2) rationale augmented CNN (RA-CNN) (Zhang et al., 2016). RA-CNN first trains a CNN-based sentence classifier to estimate the probability that a given sentence contains rationale words. Then RA-CNN scales the contribution of each sentence to the overall representation in proportion to these estimates. The final prediction is made from this overall representation.

Transfer methods We compare against two variants of our method: 1) TRANS, an attention-based classifier that does not use human rationales from the target task; 2) RA-TRANS, an attention-based classifier that directly uses human rationales to supervise the attention. Specifically, TRANS only optimizes the cross-entropy loss $\mathcal{L}^{\mathcal{T}}_{lbl}$ in the objective (Eq. (2)). For RA-TRANS, the term $\mathcal{L}^{\mathcal{T}}_{att}$ in the objective Eq. (2) is replaced by the cosine distance between human rationales and the attention generated by itself. Note that both models still have the ability to transfer, as their encoders are both initialized from $\mathbf{enc}$, which has been trained on source data and unlabeled target data.

Oracle We also report the performance of an ORACLE which shares the same architecture as ours but is supervised by the oracle attention. The oracle attention is derived from a held-out dataset with large-scale annotations for the target task (see Appendix 3 for details). This helps us analyze the contribution of our R2A approach in isolation of the inherent limitations of the target tasks.

4.3 Implementation details

We use pre-trained fastText embeddings (Bojanowski et al., 2017), a 200-dimension bi-directional LSTM (Hochreiter and Schmidhuber, 1997) for the language encoder, and a 50-dimension bi-directional LSTM for the R2A encoder. Dropout (Srivastava et al., 2014) is applied with drop rate 0.1 on the word embeddings and the last hidden layers of the classifiers. All

Source	Target	SVM	RA-SVM‡	RA-CNN‡	TRANS†	RA-TRANS‡†	OURS‡†	ORACLE†
Beer aroma+palate	Beer look	74.41	74.83	74.94	72.75	76.41	**79.53**	80.29
Beer look+palate	Beer aroma	68.57	69.23	67.55	69.92	76.45	**77.94**	78.11
Beer look+aroma	Beer palate	63.88	67.82	65.72	74.66	73.40	**75.24**	75.50

Table 3: Accuracy of transferring between aspects. Models with † use labeled data from source aspects. Models with ‡ use human rationales on the target aspect.

Source	Target	SVM	RA-SVM‡	RA-CNN‡	TRANS†	RA-TRANS‡†	OURS‡†	ORACLE†
Beer look + Beer aroma + Beer palate	Hotel location	78.65	79.09	79.28	80.42	82.10	**84.52**	85.43
	Hotel cleanliness	86.44	86.68	89.01	86.95	87.15	**90.66**	92.09
	Hotel service	85.34	86.61	87.91	87.37	86.40	**89.93**	92.42

Table 4: Accuracy of transferring between domains. Models with † use labeled data from source domains and unlabeled data from the target domain. Models with ‡ use human rationales on the target task.

parameters are optimized using Adam (Kingma and Ba, 2014). We set the initial learning rate to 0.001 and divide it by 10 once the performance on the development set plateaus. For RA-TRANS, OURS and ORACLE, we tuned $\lambda_{att}^{\mathcal{T}}$ from $\{10^2, 10^1, 10^0, 10^{-1}, 10^{-2}\}$. For domain transfer, we set $\lambda_{lm} = 0.1$, $\lambda_{wd} = 0.01$ and $\lambda_{att} = 0.01$. For aspect transfer, we adapt the same hyperparameters, but set $\lambda_{wd} = 0$ as both source tasks and the target task come from the same domain. To encourage the R2A-generated attention to be consistent with the provided rationales in aspect transfer, we augment the overall training objective of R2A in Eq. (3.3) by a consistency regularization, which is computed from the cosine distance between the R2A-generated attention and the provided rationales.

In addition, computing $\mathcal{L}_{lm}$ is both time and memory inefficient because the complexity is linear to the size of the vocabulary, which can be very large. To expedite the training, we adopt a technique proposed by Mikolov et al. (2011), which randomly splits the entire vocabulary into a predefined number of bins and minimizes the loss of the bin prediction instead of the exact token prediction. We set the bin size as 100 for our experiment.

5 Results

Aspect transfer Table 3 summarizes the results of aspect transfer on the beer review dataset. Our model (OURS) obtains substantial gains in accuracy over the baselines across all three target aspects. It closely matches the performance of ORACLE with only 0.40% absolute difference.

Model	Hotel location	Hotel cleanliness	Hotel service
OURS	**84.52**	**90.66**	**89.93**
w/o $\mathcal{L}_{wd}$	82.36	89.79	89.61
w/o $\mathcal{L}_{lm}$	82.47	90.05	89.75

Table 5: Ablation study on domain transfer from beer to hotel.

Specifically, all rationale-augmented methods (RA-SVM, RA-TRANS and OURS) outperform their rationale-free counterparts on average. This confirms the value of human rationales in the low-resource settings. We observe that the transfer baseline that directly uses rationale as augmented supervision (RA-TRANS) underperforms ORACLE by a large margin. This validates our hypothesis that human rationales and attention are different.

Domain transfer Table 4 presents the results of domain transfer using 200 training examples. We use the three aspects of the beer review data together as our source tasks while use the three aspects of hotel review data as the target. Our model (OURS) shows marked performance improvement. The error reduction over the best baseline is 15.08% on average.

We compare the learning curve in Figure 4. We observe that the performance of our model steadily improves as more annotations are provided, and the improvement over other baselines is significant and consistent.

Ablation study Table 5 presents the results of an ablation study of our model in the setting of domain transfer. As this table indicates, both the language modeling objective and the Wasserstein

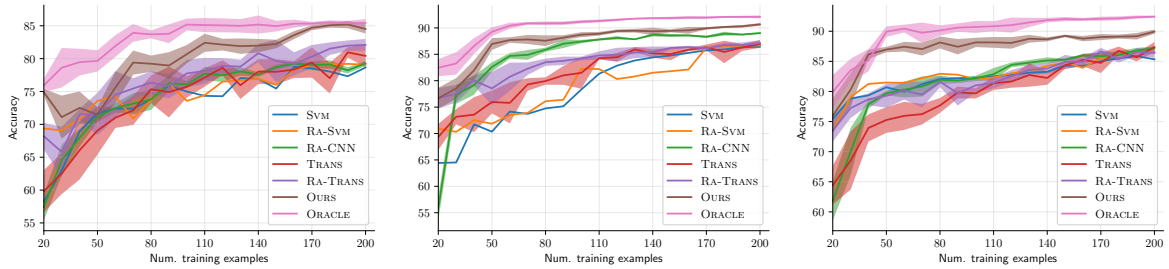
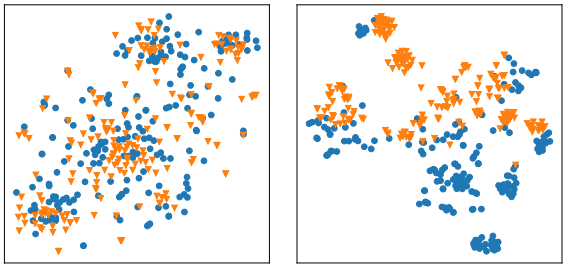


Figure 4: Learning curve of transferring from beer review to three aspects of hotel review: location (left), cleanliness (center) and service (right). For neural methods, we ran five different random seeds and plot their mean with their standard deviation.



(a) OURS (b) OURS w/o $\mathcal{L}_{wd}$

Figure 5: t-SNE visualization of the learned hidden representation[5] for beer review (blue circle) and hotel review (orange triangle).

Target	Human rationales	R2A-generated attention
Location	0.5185	**0.2371**
Cleanliness	0.5948	**0.3141**
Service	0.5833	**0.2871**

Table 6: Avg. cosine distance to the oracle attention over the target training set. The R2A is trained on beer reviews with unlabeled hotel reviews.

distance contribute similarly to the task, with the Wasserstein distance having a bigger impact.

Visualization of representation Figure 5 visualizes the hidden representation of 200 beer reviews and 200 hotel reviews using t-SNE (Maaten and Hinton, 2008). We observe that our model successfully aligns the source and the target feature distribution. This indicates the effectiveness of optimizing the Wasserstein distance objective.

Analysis of R2A-generated attention In order to validate that the trained R2A model is able to generate task-specific attention from human rationales, we perform both qualitative and quantitative analysis on the R2A-generated attention in the setting of domain transfer. It is worth pointing out that our R2A model has never seen any labeled hotel reviews during training.

Table 6 presents the average cosine distance between the R2A-generated attention and the oracle attention over the target training set. Compared with human rationales, the R2A-generated attention is much closer to the oracle attention. This explains the large performance boost of our method.

Figure 6 visualize the R2A-generated attention on the same hotel review with human rationales corresponding to three different aspects. We observe that the trained R2A model is able to produce task-specific attention scores corresponding to the provided human rationale. For example, given the rationale sentence "not the cleanest rooms but bed was clean and so was bathroom", R2A recognizes that not every token is equally important, and the attention should focus more on "clean", "cleanest", "rooms" and "bathroom".

Annotating rationales versus annotating more labeled data Providing rationales for the training data roughly doubles the annotation cost (Zaidan et al., 2007). Given the same annotation budget, a natural question is: shall we collect a few labeled examples with rationales or annotate more labeled examples? To answer this question, we vary the number of training examples in the target task. Figure 7 shows the corresponding learning curve of a classifier that is trained without rationales. The reference line represents the accuracy of our approach trained on 200 examples with rationales. We notice that in order to reach the same level of performance, the rationale-free classifier requires 800, 3100, and 1900 labeled examples on the three target tasks respectively.

[5] Since the hidden representation is a sequence of variable length, we applied t-SNE on the concatenation of the first and last element: $[h_1^{\mathbf{inv}}; h_L^{\mathbf{inv}}]$.

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.

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 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.

Task: Hotel service 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 service 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.

Figure 6: Oracle attention versus R2A-generated attention on an example of hotel review for three different tasks. Words are highlighted according to the attention scores. Human rationales are shown in bold with underlines. The oracle attention is derived from large amounts of labeled hotel reviews. The R2A is trained on labeled beer reviews with unlabeled hotel reviews.

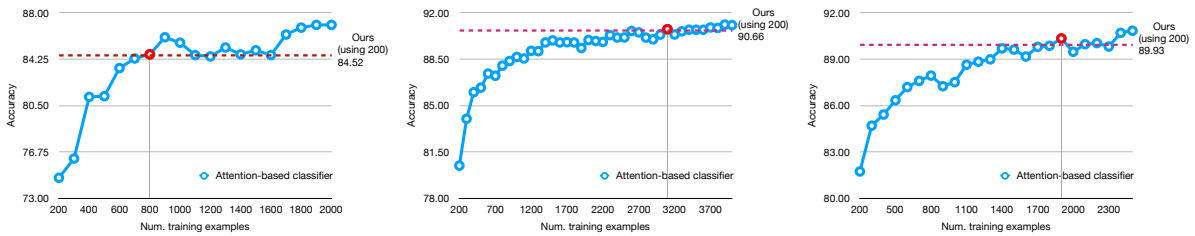


Figure 7: Learning curve of an attention-based classifier on three tasks: hotel location (left), hotel cleanliness (center), hotel service (right). The performance of our approach trained on 200 examples with human rationales is shown as a reference.

6 Conclusion

In this paper, we propose a novel approach that utilizes the connection between human rationales and machine attention to improve the performance of low-resource tasks. Specifically, we learn a transferrable mapping from rationales to high-quality attention on resource-rich tasks. The learned mapping is then used to provide an additional supervision for the target task. Experimental results on both aspect and domain transfer validate that the R2A-generated attention serves as a better form of supervision. Our model produces high-quality attention for low-resource tasks.

Acknowledgments

We thank the MIT NLP group and the reviewers for their helpful discussion and comments. This work is supported by MIT-IBM Watson AI Lab. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors, and do not necessarily reflect the views of the funding organizations.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1870–1879.

Minmin Chen, Zhixiang Xu, Kilian Weinberger, and Fei Sha. 2012. Marginalized denoising autoencoders for domain adaptation. *arXiv preprint arXiv:1206.4683*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 513–520.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5769–5779.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Hidetaka Kamigaito, Katsuhiko Hayashi, Tsutomu Hirao, Hiroya Takamura, Manabu Okumura, and Masaaki Nagata. 2017. Supervised attention for sequence-to-sequence constituency parsing. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 7–12.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Neural machine translation with supervised attention. *arXiv preprint arXiv:1609.04186*.

Shulin Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2017. Exploiting argument information to improve event detection via supervised attention mechanisms. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1789–1798.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1020–1025. IEEE.

Tomáš Mikolov, Stefan Kombrink, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2011. Extensions of recurrent neural network language model. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5528–5531. IEEE.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 783–792. ACm.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using annotator rationales to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267.

Ye Zhang, Iain Marshall, and Byron C Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 795. NIH Public Access.

Yuan Zhang, Regina Barzilay, and Tommi Jaakkola. 2017. Aspect-augmented adversarial networks for domain adaptation. *Transactions of the Association of Computational Linguistics*, 5(1):515–528.

Guangyou Zhou, Zhiwen Xie, Jimmy Xiangji Huang, and Tingting He. 2016. Bi-transferring deep neural networks for domain adaptation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 322–332.