Rationalizing Neural Predictions

Tao Lei, Regina Barzilay, Tommi Jaakkola

Guillaume Bressan
Théo Delemazure
Jean Dupin

Why do we need to explain predictions? A not so NLP example

Why do we need to explain predictions? A not so NLP example

Now let's look at **which part of the images** had the biggest impact on the final decision ...

Why do we need to explain predictions? A not so NLP example

It is actually a perfect **snow detector** !

Why do we need to explain predictions? A not so NLP example

In many applications, prediction from neural network are used to drive **critical decisions**.

We need to understand, and most importantly, to **explain these decisions**.

Why do we need to explain predictions? Let's go back to NLP

The beer wasn't what I expected, and I'm not sure it's "true to style", but I thought it was delicious. A very pleasant reby red-amber color with a relatively brilliant finish, but a limited amount of carbonation, from the look of it; Aroma is what I think an amber ale should be - a nice blend of caramel and happiness bound together.

"Look rating" prediction : 5/5

Which sentence of the review would you use **to explain the prediction** ?

It should be **concise** and **relevant**

Why do we need to explain predictions? Let's go back to NLP

The beer wasn't what I expected, and I'm not sure it's "true to style", but I thought it was delicious. ***A very pleasant reby red-amber color*** *with a relatively brilliant finish, but a limited amount of carbonation, from the look of it; Aroma is what I think an amber ale should be - a nice blend of caramel and happiness bound together.*

"Look rating" prediction : 5/5

Which sentence of the review would you use **to explain the prediction** ?

It should be **concise** and **relevant**

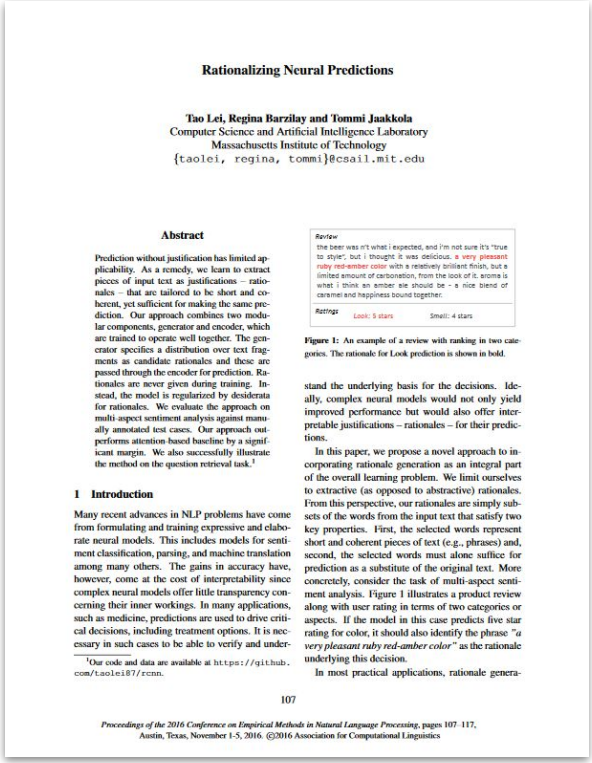
“Rationalizing Neural Prediction”

Authors : Tao Lei, Regina Barzilay and Tommi Jaakkola (*MIT*)

Year of publication : 2016

Published in : *Conference on Empirical Methods in Natural Language Processing*

Scientific impact : 263 citations to this day



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

Abstract

Prediction without justification has limited applicability. As a remedy, we learn to extract pieces of input text as justifications – rationales – that are tailored to be short and coherent, yet sufficient for making the same prediction. Our approach combines two modular components, generator and encoder, which are trained to operate well together. The generator specifies a distribution over text fragments as candidate rationales and these are passed through the encoder for prediction. Rationales are never given during training. Instead, the model is regularized by desiderata for rationales. We evaluate the approach on multi-aspect sentiment analysis against manually annotated test cases. Our approach outperforms attention-based baseline by a significant margin. We also successfully illustrate the method on the question retrieval task.[1]

Review
the beer was n't what i expected, and i'm not sure it's "true to style", but i thought it was delicious. a very pleasant ruby red-amber color with a relatively brilliant finish, but a limited amount of carbonation, from the look of it. aroma is what i think an amber ale should be - a nice blend of caramel and happiness bound together.

Ratings Look: 5 stars Smell: 4 stars

Figure 1: An example of a review with ranking in two categories. The rationale for Look prediction is shown in bold.

1 Introduction

Many recent advances in NLP problems have come from formulating and training expressive and elaborate neural models. This includes models for sentiment classification, parsing, and machine translation among many others. The gains in accuracy have, however, come at the cost of interpretability since complex neural models offer little transparency concerning their inner workings. In many applications, such as medicine, predictions are used to drive critical decisions, including treatment options. It is necessary in such cases to be able to verify and understand the underlying basis for the decisions. Ideally, complex neural models would not only yield improved performance but would also offer interpretable justifications – rationales – for their predictions.

In this paper, we propose a novel approach to incorporating rationale generation as an integral part of the overall learning problem. We limit ourselves to extractive (as opposed to abstractive) rationales. From this perspective, our rationales are simply subsets of the words from the input text that satisfy two key properties. First, the selected words represent short and coherent pieces of text (e.g., phrases) and, second, the selected words must alone suffice for prediction as a substitute of the original text. More concretely, consider the task of multi-aspect sentiment analysis. Figure 1 illustrates a product review along with user rating in terms of two categories or aspects. If the model in this case predicts five star rating for color, it should also identify the phrase *"a very pleasant ruby red-amber color"* as the rationale underlying this decision.

In most practical applications, rationale genera-

[1] Our code and data are available at https://github.com/taolei87/rcnn.

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 107–117,
Austin, Texas, November 1-5, 2016.

Outline

I **Introduction to the problem**

II **The paper**

1. The method used
2. The experiments

III **Our comments**

Learning rationales

What is a rationale ?

- Think of a review as a **vector of words**

$$x = (\text{Very, dark, beer, Pours, a, nice, ...})$$

What is a rationale ?

- Think of a review as a **vector of words**

$$x = (\text{Very, dark, beer, Pours, a, nice, ...})$$

- **A rationale** is a small subset of words from the original review which ideally have the same predictive power as the full review for inferring ratings. You can think of it as a binary vector z (with $z_i = 1$ iff the i^{th} word is selected).

$$z = (1, 1, 1, 0, 0, 0, ...)$$

What is a rationale ?

- Think of a review as a **vector of words**

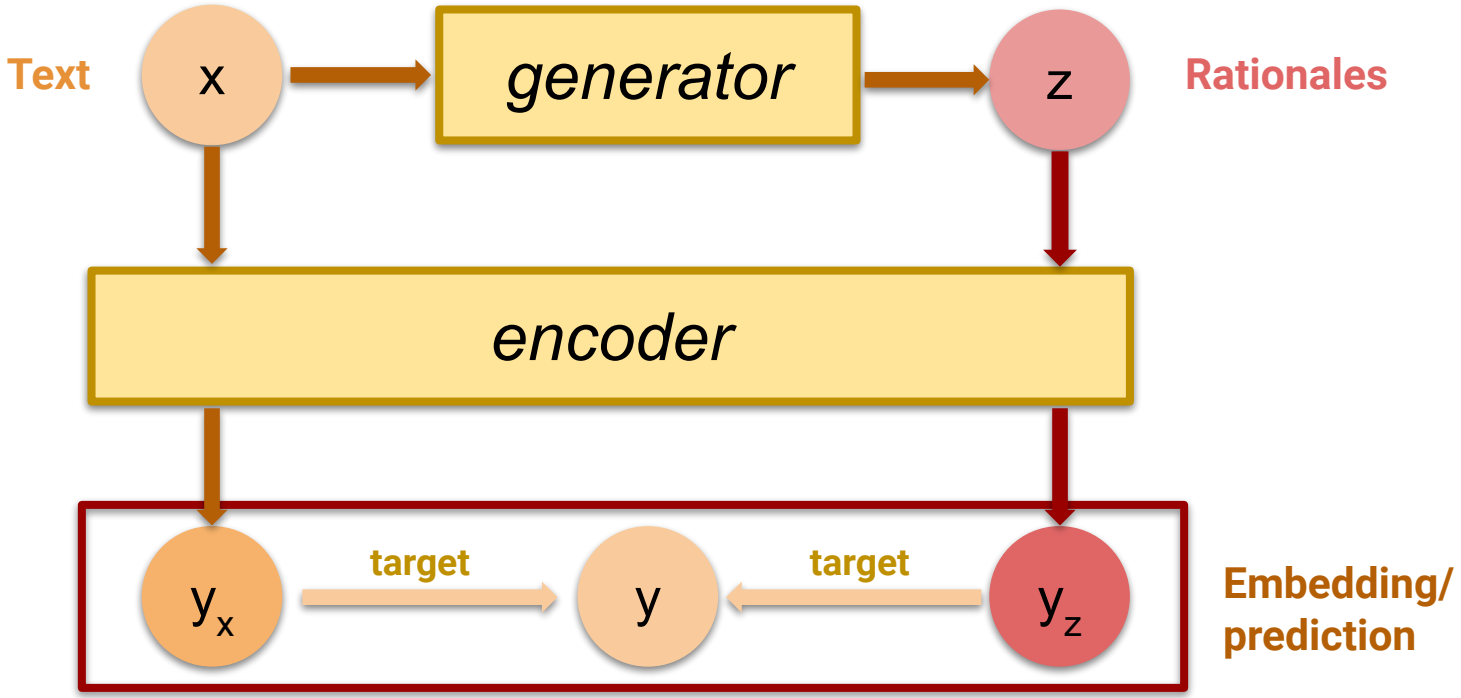
$$x = (\text{Very, dark, beer, Pours, a, nice, ...})$$

- **A rationale** is a small subset of words from the original review which ideally have the same predictive power as the full review for inferring ratings. You can think of it as a binary vector z (with $z_i = 1$ iff the i^{th} word is selected).

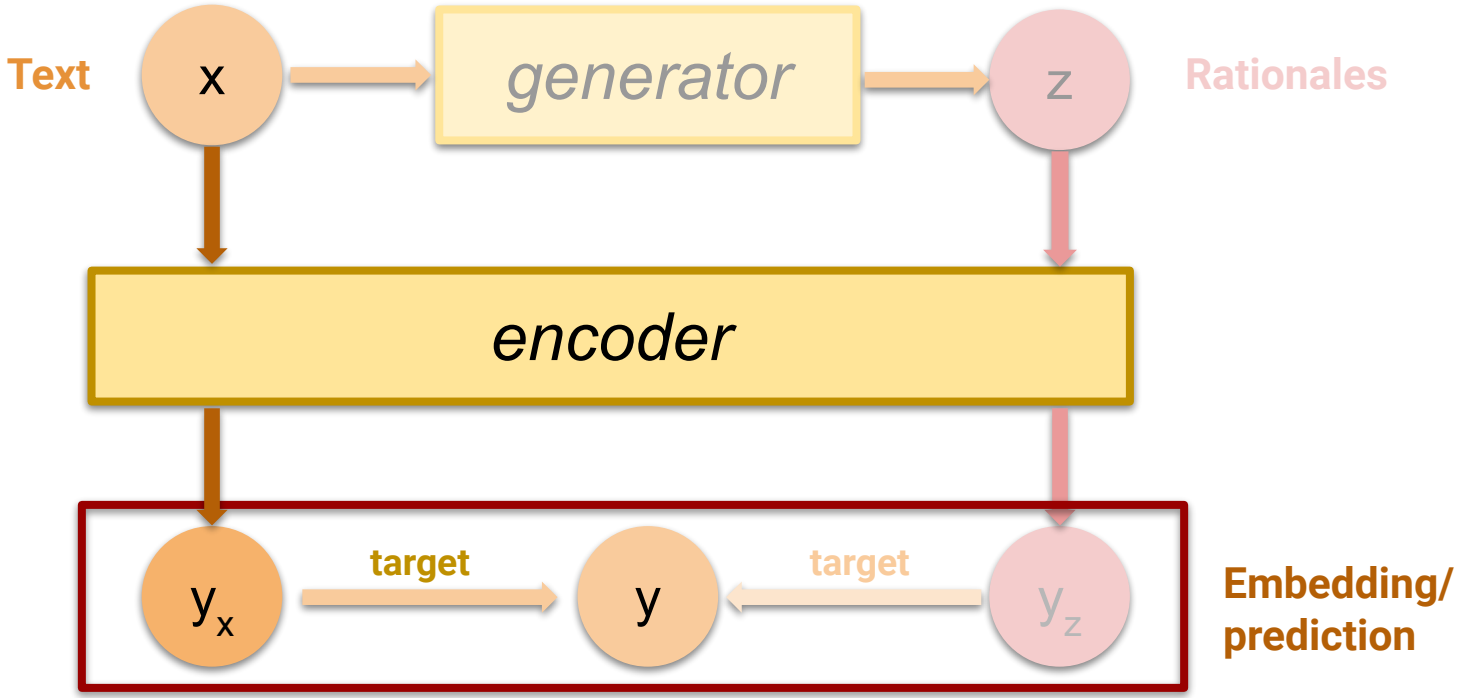
$$z = (1, 1, 1, 0, 0, 0, ...)$$

Can we learn to “generate” such a rationale in a completely **unsupervised** fashion?

How to learn rationales ?



The encoder


Text
x
generator
z
Rationales
encoder
y_x
target
y
target
y_z
Embedding/
prediction

The encoder

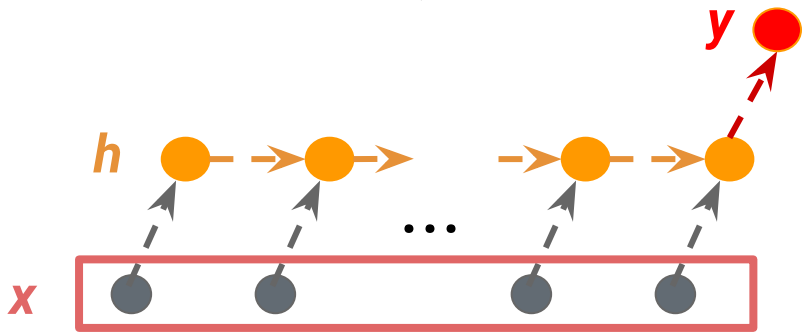
Depend on the task, for a regression task for instance

Loss function

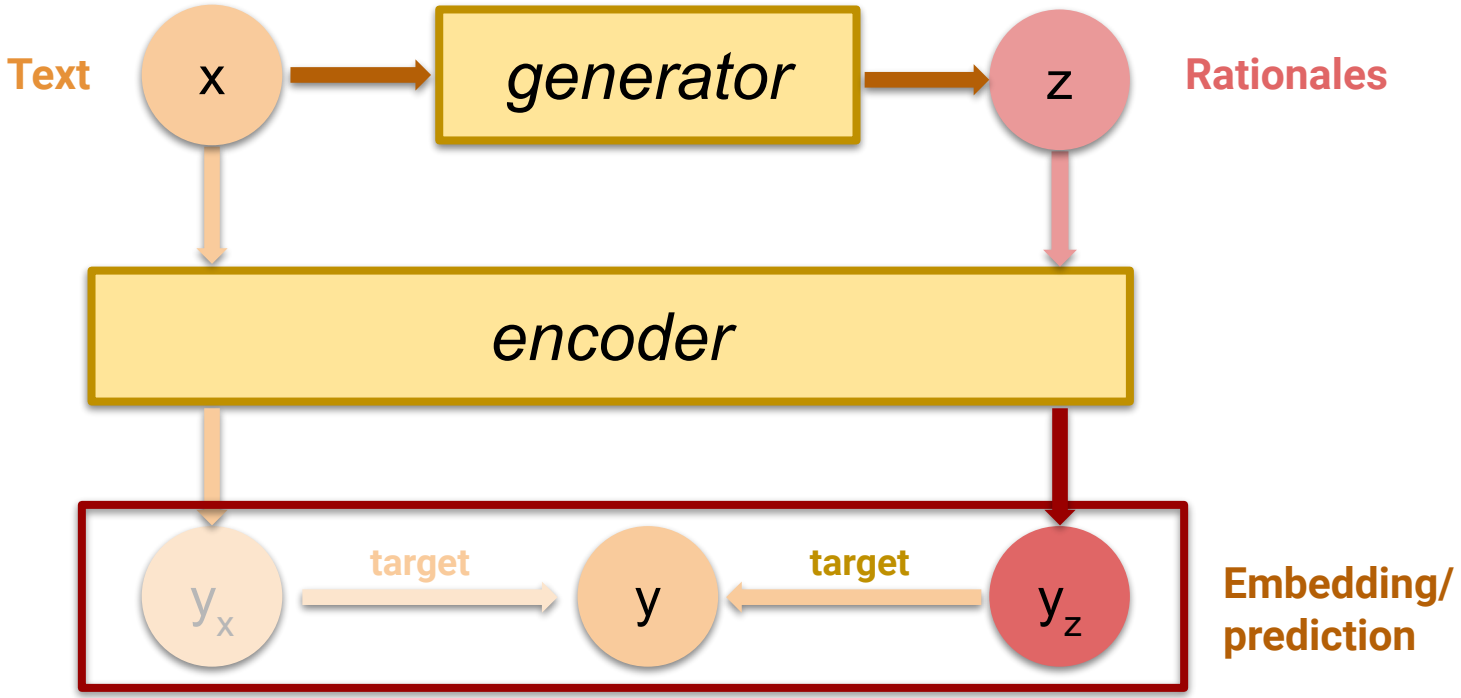
$$\mathcal{L}(x, y) = \|enc(x) - y\|$$

Network

RNN, LSTM...



The generator



The generator : Loss on the rationales

1. Rationale ***z*** produced must suffice as a replacement for the input text ***x*** for prediction i.e inference made solely on rationale **must be "close"** to target sentiment vector. For instance in a regression task :

$$\mathcal{L}(z, x, y) = \|enc(z, x) - y\|$$

The generator : Loss on the rationales

1. Rationale ***z*** produced must suffice as a replacement for the input text ***x*** for prediction i.e inference made solely on rationale **must be "close"** to target sentiment vector. For instance in a regression task :

$$\mathcal{L}(z, x, y) = \|enc(z, x) - y\|$$

2. We should select **few words** and selections should form phrases (i.e. **consecutive words**) -> regularizer

$$\Omega(z) = \lambda_1 ||z|| + \lambda_2 \sum_t |z_t - z_{t-1}|$$

The generator : Loss on the rationales

1. Rationale ***z*** produced must suffice as a replacement for the input text ***x*** for prediction i.e inference made solely on rationale **must be "close"** to target sentiment vector. For instance in a regression task :

$$\mathcal{L}(z, x, y) = \|enc(z, x) - y\|$$

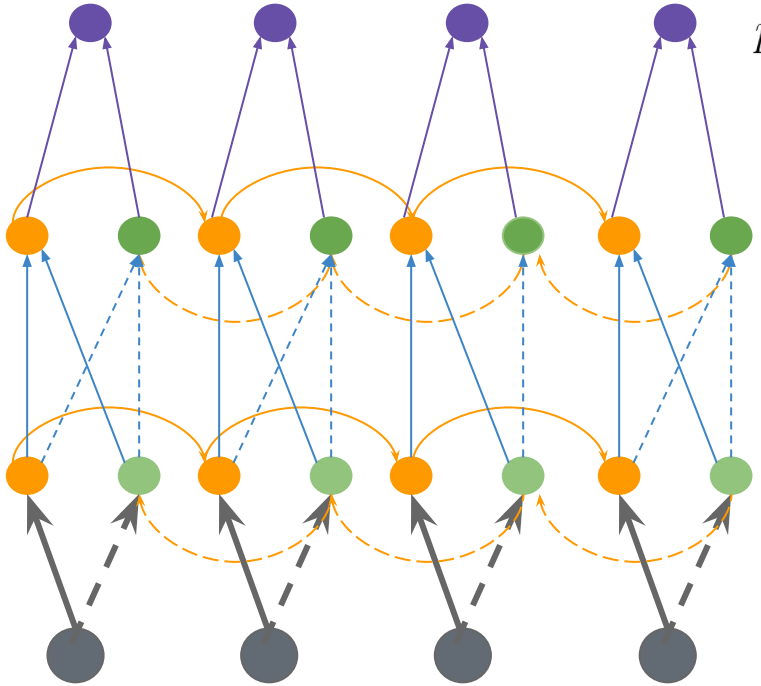
2. We should select **few words** and selections should form phrases (i.e. **consecutive words**) -> regularizer

$$\Omega(z) = \lambda_1 ||z|| + \lambda_2 \sum_t |z_t - z_{t-1}|$$

3. Loss to minimise over training examples x and y

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

Choice of architecture



$$p(z_t|x) = \sigma_z(W^z[\overrightarrow{h_t}; \overleftarrow{h_t}] + b^z)$$

$$p(z|x) = \prod_{t=1}^{l} p(z_t|x)$$

Independent selection

$$\overleftarrow{h_t} = \overleftarrow{f}(x_t, \overleftarrow{h_{t+1}})$$

$$\overrightarrow{h_t} = \overrightarrow{f}(x_t, \overrightarrow{h_{t-1}})$$

$$x = (x_1, ..., x_l)$$

What is *f*?

	D	d	l	$\|\theta\|$	MSE
SVM	260k	-	-	2.5M	0.0154
SVM	1580k	-	-	7.3M	0.0100
LSTM	260k	200	2	644k	0.0094
RCNN	260k	200	2	323k	**0.0087**

Table 3: Comparing neural encoders with bigram SVM model. MSE is the mean squared error on the test set. D is the amount of data used for training and development. d stands for the hidden dimension, l denotes the depth of network and $|\theta|$ denotes the number of parameters (number of features for SVM).

Doubly stochastic gradient

In practice, ***gen(x)*** is not a rationales but a probability distribution over the rationales, and we have

$$z \sim gen(x) \equiv \mathbb{P}(z|x)$$

How can we compute **the gradient of** $\mathbb{E}_{z\sim gen(x)}[cost(z,x,y)]$ **?**

Doubly stochastic gradient

In practice, ***gen(x)*** is not a rationales but a probability distribution over the rationales, and we have

$$z \sim gen(x) \equiv \mathbb{P}(z|x)$$

How can we compute **the gradient of** $\mathbb{E}_{z\sim gen(x)}[cost(z,x,y)]$ **?**

Doubly stochastic gradient !

Doubly stochastic gradient

$$\frac{\partial \mathbb{E}_{z\sim gen(x)}[cost(z,x,y)]}{\partial \theta_{gen}} = \mathbb{E}_{z\sim gen(x)}[cost(z,x,y)\frac{\partial log(\mathbb{P}(z|x))}{\partial \theta_{gen}}]$$

$$\frac{\partial \mathbb{E}_{z\sim gen(x)}[cost(z,x,y)]}{\partial \theta_{enc}} = \mathbb{E}_{z\sim gen(x)}[\frac{\partial cost(z,x,y)}{\partial \theta_{enc}}]$$

Just sample a bunch of rationales and approximate

Experimental results

Beer reviews : Experimental setup

The dataset:

- *BeerAdvocate* review dataset (**1.5m** reviews)
- In addition, dataset provides **ratings** (on scale of 0 to 5) for each **aspect(appearance, smell, palate, taste)** as well as overall rating
- **Sentence-level annotations** on around **1000** reviews -> ***test set***

Beer reviews : Experimental setup

The dataset:

- *BeerAdvocate* review dataset (**1.5m** reviews)
- In addition, dataset provides **ratings** (on scale of 0 to 5) for each **aspect(appearance, smell, palate, taste)** as well as overall rating
- **Sentence-level annotations** on around **1000** reviews -> ***test set***

Pre-processing:

- Correlation between any pair of aspects ratings (e.g. taste vs smell) or any aspect vs overall rating (e.g. taste vs overall score) is high at **63.5%**
- If trained directly on the dataset as is, run the risk of not properly isolating the various aspects **-> pre-processing by picking "less correlated"** examples

Beer reviews : Results

Very dark beer. Pours a nice finger and a half of creamy foam and stays throughout the beer. Smells of coffee and roasted malt . Has a major coffee-like taste with hints of chocolate. If you like black coffee , you will love this porter. Creamy smooth mouthfeel and definitely gets smoother on the palate once it warms. It's an ok porter but i feel there are much better one 's out there .

Red = appearance
Blue = smell
Green = palate

Precision metric: how many of the selected words are in the annotated sentences from the test set?

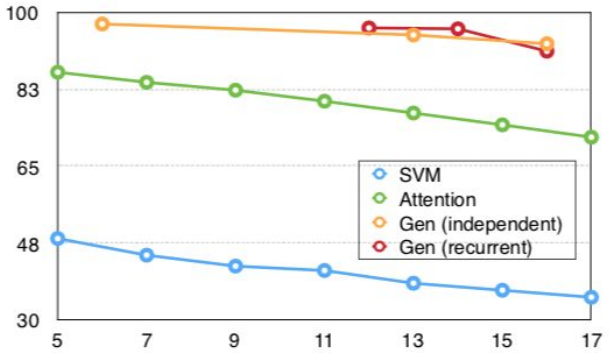


Figure 4: Precision (y-axis) when various percentages of text are extracted as rationales (x-axis) for the appearance aspect.

Beer reviews : Results

Method	Appearance		Smell		Palate	
	% precision	% selected	% precision	% selected	% precision	% selected
SVM	38.3	13	21.6	7	24.9	7
Attention model	80.6	13	88.4	7	65.3	7
Generator (independent)	94.8	13	93.8	7	79.3	7
Generator (recurrent)	96.3	14	95.1	7	80.2	7

Table 2: Precision of selected rationales for the first three aspects. The precision is evaluated based on whether the selected words are in the sentences describing the target aspect, based on the sentence-level annotations. Best training epochs are selected based on the objective value on the development set (no sentence annotation is used).

AskUbuntu questions : Experimental setup

The dataset:

- *AskUnbuntu* dataset : 167k **unique questions** (each consisting a question title and a body)
- Instead of ratings, we have 16k user-identified similar **question pairs**. The model is trained with each question along with 20 others, each couple being labeled 0 or 1 following similarity
- 400×20 **query-candidate question pairs** are annotated for evaluation.

What should we compare the results with ?

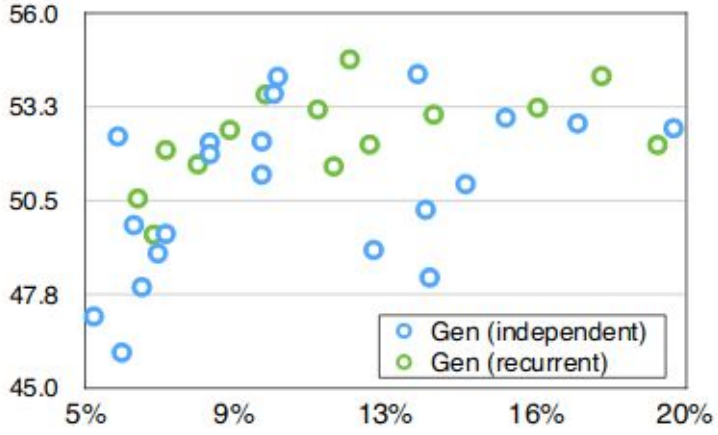
- We are already granted information extraction : **titles** (and bodies somehow)

AskUbuntu questions : Results

how do i mount a hibernated partition with windows 8 in ubuntu ? i ca n't mount my other partition with windows 8 , i have ubuntu 12.10 amd64 : error mounting /dev/sda1 at <unk> : command-line `mount -t `` ntfs '' -o `` uhelper=udisks2 , nodev , nosuid , uid=1000 , gid=1000 , dmask=0077 , fmask=0177 '' `` /dev/sda1 '' `` <unk> '' ' exited with non-zero exit status 14 : windows is hibernated , refused to mount . failed to mount '/dev/sda1 ' : operation not permitted the ntfs partition is hibernated . please resume and shutdown windows properly , or mount the volume read-only with the 'ro ' mount option

	MAP (dev)	MAP (test)	%words
Full title	56.5	60.0	10.1
Full body	54.2	53.0	89.9
Independent	55.7	53.6	9.7
	56.3	52.6	19.7
Dependent	56.1	54.6	11.6
	56.5	55.6	32.8

Table 4: Comparison between rationale models (middle and bottom rows) and the baselines using full title or body (top row).



Personal comments

Comments

- **Training**
  - **Approximation of expectation** in the loss function to minimise requires sampling. Authors don't indicate their approach to choosing **appropriate number of samples** (and therefore avoid mentioning the approximation error to the expectation)
- **Inference**
  - Once the generator probability distribution has been fully trained, how is a specific rationale **z chosen over all possibilities?** Threshold, argmax of probability distribution?
- **Construction of loss function**
  - Why choose L2-norm to minimise reconstruction error? Wouldn't L1-norm be more appropriate in that context to put an emphasis on sparsity?

Scientific impact

Recent research quoting this paper

- This paper follow the trend on **Interpretability** in machine learning. A lot of **recent surveys** on the subject are quoting this paper:
    - *"Towards A Rigorous Science of Interpretable Machine Learning"* (F. Doshi-Velez and B. Kim)
    - *"A Survey of Methods for Explaining Black Box Models"* (Guidotti et Al.)
- Other **explainability methods** are introduced and compared to this one.

Thank you !