A causal framework for explaining the predictions of black-box sequence-to-sequence models

David Alvarez-Melis and Tommi S. Jaakkola
Presenter: Ji Gao

https://qdata.github.io/deep2Read

Outline

A causal framework for explaining the predictions of black-box sequence-to-sequence models
EMNLP'17

- Black-box intrepretation on NLP sequence generation tasks.
- Explanation: A sets of input and output tokens that have causal dependencies under the model.
- Adopt a VAE to generate semantically related sentence variations

Related Work

- Local Interpretable Model-agnostic Explanations (LIME)[RSG16]

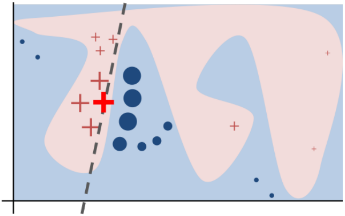


Figure 3: Toy example to present intuition for LIME.

Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K
$\mathcal{Z} \leftarrow \{\}$
for $i \in \{1, 2, 3, ..., N\}$ **do**
$z'_i \leftarrow sample\_around(x')$
$\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
end for
$w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ $\triangleright$ with z'_i as features, $f(z)$ as target
return w

- Local Interpretable Model-agnostic Explanations (LIME)[RSG16]
- On classification task. Other works include [LBJ16]

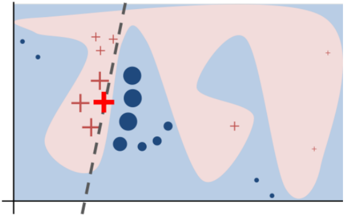


Figure 3: Toy example to present intuition for LIME.

Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K

```
Z ← {}
for i ∈ {1, 2, 3, ..., N} do
    z'_i ← sample_around(x')
    Z ← Z ∪ ⟨z'_i, f(z_i), π_x(z_i)⟩
end for
w ← K-Lasso(Z, K)   ▷ with z'_i as features, f(z) as target
return w
```

Definition

Model

A black-box model is defined as $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$.
Input $\mathbf{x} \in \mathcal{X} = \{x_1, x_2, .., x_n\}$, output $\mathbf{y} \in \mathcal{Y} = \{y_1, y_2, .., y_n\}$

Assumption

The behaviour of the model can be represented as a bipartite graph $G = (V_x \cup V_y, E)$.
V_x and V_y are elements in $\mathbf{x}$ and $\mathbf{y}$, respectively.
An edge E_{ij} is weighted with the occurrence of token x_i and y_j.

Definition

Assumption

The behaviour of the model can be represented as a bipartite graph $G = (V_x \cup V_y, E)$.
V_x and V_y are elements in $\mathbf{x}$ and $\mathbf{y}$, respectively.
An edge E_{ij} is weighted with the occurrence of token x_i and y_j.

Explanation

An explanation is a collection of sub-graphs in G.
Suppose a component $G^k = (V_x^k \cup V_y^k, E^k)$, then an explanation $E_{x \to y} = \{G^1, ..., G^k\}$

Pipeline

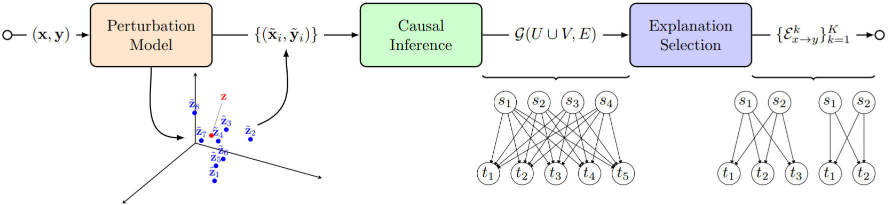


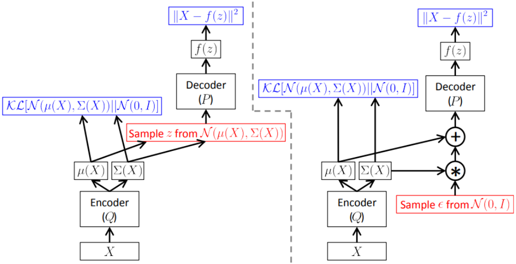
Figure 1: A schematic representation of the proposed prediction interpretability method.

3 steps:

1. Generate perturbed versions of inputs
2. Use the perturbed inputs to estimate a causal graph model
3. Generate explanations(Subgraphs)

Step 1: Perturbation Model

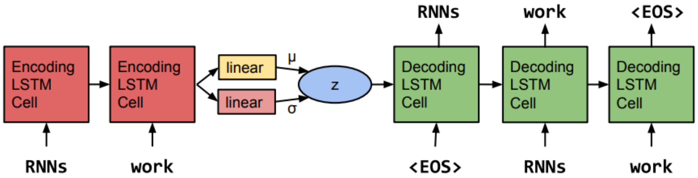
- Generate perturbation on arbitrary structured data is difficult.
- Use a VAE to generate perturbation: sample small perturbation on latent space, and use the decoder to generate perturbed samples.
- Discrete VAE model from [BVV+15]
- Scale the variance to get different levels of perturbation



VAE on text sequence[BVV+15]

- Inspired from Variational Recurrent Autoencoder(VRAE)
- **Key equation:**

$$L(\theta; x) = -KL(q_\theta(z|x)||p(z)) + \mathbb{E}_{q_\theta(z|x)}[\log p_\theta(x|z)]$$



Samples generated by VAE model

Sampling temperature α ↓			
	Input:	Students said they looked forward to his class .	The part you play in making the news is very important .
	Perturbations	Students said they looked forward to his class	The part with play in making the news is important .
		Students said they looked forward to his history .	The question you play in making the funding is a important .
		Students said they looked around to his class .	The part was created in making the news is very important .
		Some students said they really went to his class .	This part you play a place on it is very important .
		Students know they looked forward to his meal .	The one you play in making the news is very important .
		Students said they can go to that class .	These part also making newcomers taken at news is very important .
		You felt they looked forward to that class .	The terms you play in making the news is very important .
		Producers said they looked forward to his cities .	This part made play in making the band , is obvious .
		Note said they looked forward to his class .	The key you play in making the news is very important .
		Students said they tried thanks to the class ;	The part respect plans in making the pertinent survey is available .
		Why they said they looked out to his period .	In part were play in making the judgment , also important .
		Students said attended navigate to work as deep .	The issue met internationally in making the news is very important .
		What having they : visit to his language ?	In 50 interviews established in place the news is also important .
		Transition said they looked around the sense ."	The part to play in making and safe decision-making is necessary .
		What said they can miss them as too .	The order you play an making to not still unique .

Table 3: Samples generated by the English VAE perturbation model around two example input sentences for increasing scaling parameter α.

Causal model

- Use logistic regression to estimate the model

$$P(y_j \in \tilde{y} | \tilde{x}) = \sigma(\theta_j^T \phi_x(\tilde{x})) \quad (1)$$

$\phi_x \in \{0, 1\}^{|x|}$ is the binary embedding vector of sample **x**.

Explanation selection

- Model the graph partitioning problem into a MIP programming problem[FZP12]

$$\min_{(x_{ik}^u, x_{jk}^v, y_{ij}) \in Y} \sum_{i=1}^{n} \sum_{j=1}^{m} \theta_{ij} y_{ij} + \max_{\substack{S: S \subset J, |S| \leq \Gamma \\ (i_t, j_t) \in J/S}} \sum_{(i,j) \in S} \hat{\theta}_{ij} y_{ij} + (\Gamma - |\Gamma|) \hat{\theta}_{i_t, j_t} y_{i_t, j_t} \tag{2}$$

- After solving the problem, sort the importance defined as importance$(E^k) = -\sum_{(i,j) \in X_k} \theta_{ij}$.
- Return the top ranked slices.

Algorithm

Algorithm 1 Structured-output causal rationalizer

procedure SOCRAT($\mathbf{x}, \mathbf{y}, F$)
  $(\boldsymbol{\mu}, \boldsymbol{\sigma}) \leftarrow$ ENCODE($\mathbf{x}$)
  for $i = 1$ **to** N **do**
    $\tilde{\mathbf{z}}_i \leftarrow$ SAMPLE($\boldsymbol{\mu}, \boldsymbol{\sigma}$)
    $\tilde{\mathbf{x}}_i \leftarrow$ DECODE($\tilde{\mathbf{z}}_i$)
    $\tilde{\mathbf{y}}_i \leftarrow F(\tilde{\mathbf{x}}_i)$
  end for
  $G \leftarrow$ CAUSAL($\mathbf{x}, \mathbf{y}, \{\tilde{\mathbf{x}}_i, \tilde{\mathbf{y}}_i\}_{i=1}^{N}$)
  $E_{x \mapsto y} \leftarrow$ BIPARTITION(G)
  $E_{x \mapsto y} \leftarrow$ SORT($E_{x \mapsto y}$) ▷ By cut capacity
  return $E_{x \mapsto y}$
end procedure

Perturbation Model.

Model Details in experiments

- For VAE to handle sequential input, use stacked RNN on both sides. and a stacked variational layer.
- Use optimization library gurobi to solve the partition models at a MIP problem

Experiment 1: Recovering simple mapping

- Dataset: CMU dictionary of word pronunciation, mapping words to phonemes. Including 130K words.
- *vowels* $\rightarrow$ $V\ AW1\ AH0\ L\ Z$
- Result:

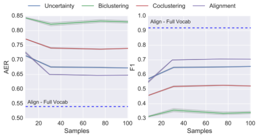


Figure 2: Arpabet test results as a function of number of perturbations used. Shown are mean plus confidence bounds over 5 repetitions. **Left**: Alignment Error Rate, **Right**: F1 over edge prediction.

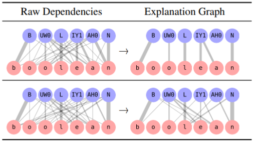


Table 1: Inferred dependency graphs before (left) and after (right) explanation selection for the prediction: *boolean* $\mapsto$ `B UW0 L IY1 AH0 N`, in independent runs with large (top) and small (bottom) clustering parameter k.

Machine Translation

- Use multiple models: Azure, NMT, Human

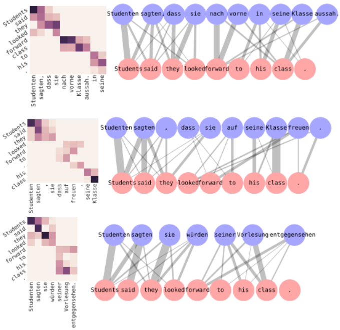


Figure 3: Explanations for the predictions of three Black-Box translators: Azure (top), NMT (middle) and human (bottom). Note that the rows and columns of the heatmaps are permuted to show explanation *chunks* (clusters).

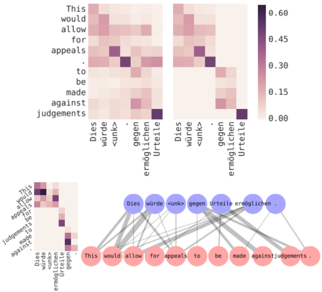


Figure 4: **Top**: Original and clustered attention matrix of the NMT system for a given translation. **Bottom**: Dependency estimates and explanation graph generated by SocRat with with $S = 100$.

"mediocre" Dialogue System

- Use a Seq2seq model on OpenSubtitle corpus

Input	Prediction
What do you mean it doesn't matter?	*I don't know*
Perhaps have we met before?	*I don't think so*
Can I get you two a cocktail?	*No, thanks.*

Table 2: "Good" dialogue system predictions.

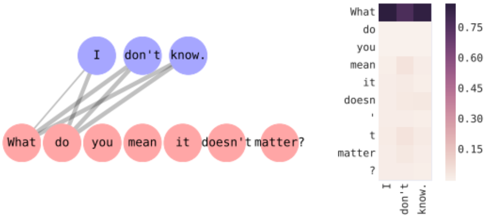


Figure 5: Explanation with $S = 50$ (left) and attention (right) for the first prediction in Table 2.

Bias detection

- Simulate a biased corpus: In a English to French dataset, prepend the word 'However' when the translation includes every informal registry(e.g. *tu*)

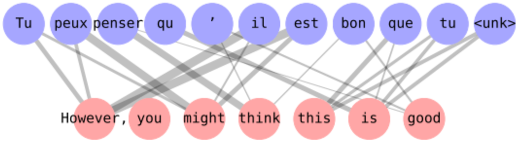
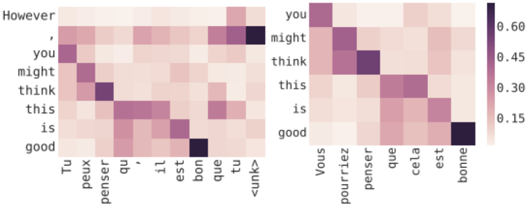


Figure 6: Explanation with $S = 50$ for the prediction of the biased translator.



Bias detection

- Use Azure model, translate several simple French sentences that lacks gender specification in English, but require gender-declined words in the output.

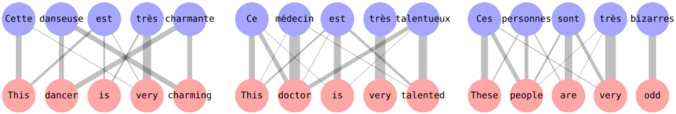


Figure 8: Explanations for biased translations of similar gender-neutral English sentences into French generated with Azure's MT service. The first two require gender declination in the target (French) language, while the third one, in plural, does not. The dependencies in the first two shed light on the cause of the biased selection of gender in the output sentence.

Reference I

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio, *Generating sentences from a continuous space*, arXiv preprint arXiv:1511.06349 (2015).

Neng Fan, Qipeng P Zheng, and Panos M Pardalos, *Robust optimization of graph partitioning involving interval uncertainty*, Theoretical Computer Science **447** (2012), 53–61.

Tao Lei, Regina Barzilay, and Tommi Jaakkola, *Rationalizing neural predictions*, arXiv preprint arXiv:1606.04155 (2016).

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, *Why should i trust you?: Explaining the predictions of any classifier*, Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, ACM, 2016, pp. 1135–1144.