Generating Hierarchical Explanations on Text Classification via Feature Interaction Detection
基于特征交互方法在文本分类任务中生成层次的可解释性

修晔良

Motivation:

现有方法从输入文本中选择单词或短语作为解释，但是忽 略了它们之间的相互作用。

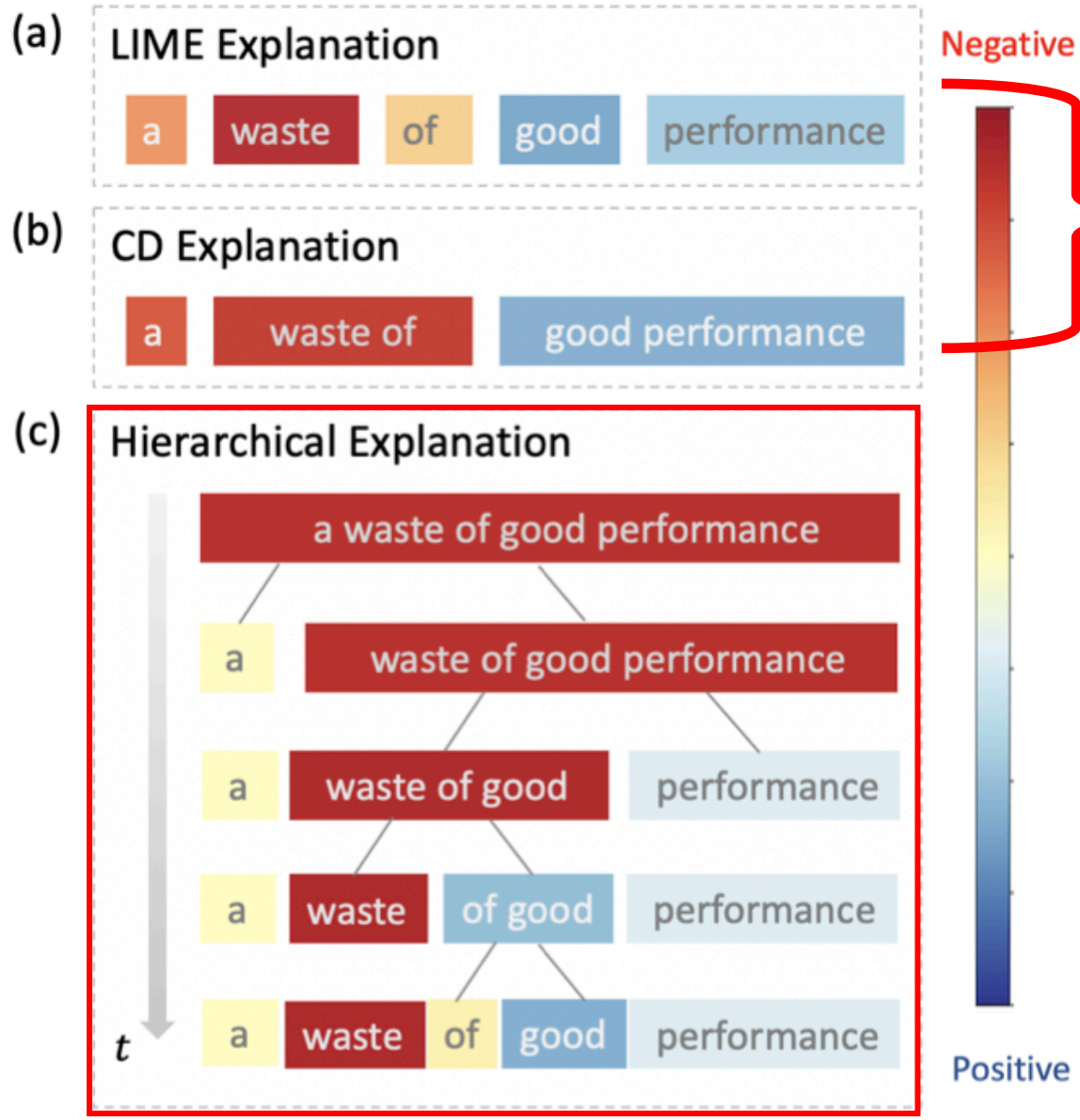
Contribution:

- 通过检测特征交互来构建层次性的解释。 这种解释可视化了单词和短语如何在层次结构中在不同级别上进行组合，这可以帮助用户理解黑盒模型的决策。
- 通过自动和人工评估，在两个基准数据集上使用三个文本分类器（LSTM，CNN和BERT）对所提出的方法进行了评估。

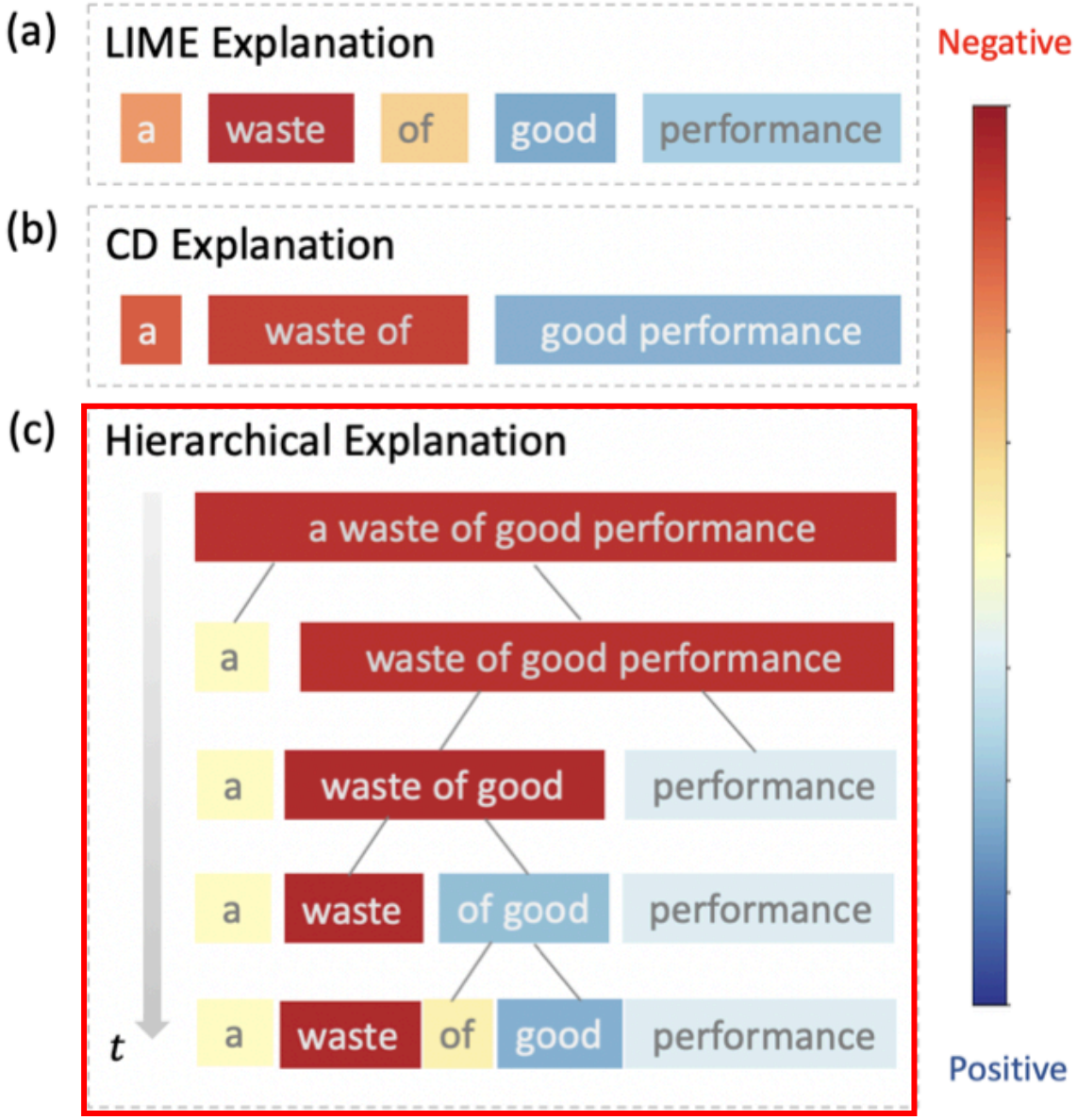
单词级的解释

短语级的解释

通过不同粒度的特征（例如单词或短语）交互来揭示其对样本预测的贡献度



不能够根据单词和短语如何相互作用以及如何共同构成最终预测来解释模型决策



Question

- How to build hierarchical explanations?
- How to detect the feature interaction?
- How to quantify feature importance ?

Generating Hierarchical Explanations

For the next timestep, which text span the algorithm should pick to split and where is the dividing point?

$x = (x_1, x_2, \ldots, x_n)$ ← A waste of good performance

$P = \{x_{(0,s_1]}, x_{(s_1,s_2]}, \ldots, x_{(p-1,n]}\}$ ← A waste of good performance

The Interaction Score，The inner回答了给定一个文本块，从哪个点进行划分的问题

$$\min_{\boldsymbol{x}_{(s_i,s_{i+1}]} \in \mathcal{P}} \min_{j \in (s_i, s_{i+1})} \phi(\boldsymbol{x}_{(s_i,j]}, \boldsymbol{x}_{(j,s_{i+1}]} \mid \mathcal{P}), \quad (1)$$

Detecting Feature Interaction- Shapley Interaction Index

$$\phi(j_1,j_2\,|\mathcal{P}) = \sum_{S\subseteq\mathcal{N}\setminus\{j_1,j_2\}} \frac{|S|!(P-|S|-1)!}{P!}\gamma(j_1,j_2,S), \tag{2}$$

$$\gamma(j_1,j_2,S) = \mathbb{E}[f(\boldsymbol{x}')\,|\,S\cup\{j_1,j_2\}] - \mathbb{E}[f(\boldsymbol{x}')\,|\,S\cup\{j_1\}] - \mathbb{E}[f(\boldsymbol{x}')\,|\,S\cup\{j_2\}] + \mathbb{E}[f(\boldsymbol{x}')\,|\,S], \tag{3}$$

New Partition

$$\mathcal{N} = \mathcal{P}\setminus\{\boldsymbol{x}_{(s_i,s_{i+1}]}\}\cup\{\boldsymbol{x}_{(s_i,j]},\boldsymbol{x}_{(j,s_{i+1}]}\} = \{\boldsymbol{x}_{(0,s_1]},\ldots,\boldsymbol{x}_{(s_i,j]},\boldsymbol{x}_{(j,s_{i+1}]},\ldots,\boldsymbol{x}_{(s_{P-1},n]}\}.$$

$$\phi(j_1,j_2\,|\mathcal{P}) = \sum_{S\subseteq\mathcal{N}_m\setminus\{j_1 j_2\}} \frac{|S|!(M-|S|-2)!}{(M-1)!}\gamma(j_1,j_2,S), \tag{4}$$

Quantifying Feature Importance

To measure the contribution of a feature $x(s_i, s_{i+1})$ to the model prediction

$$\psi(\boldsymbol{x}_{(s_i,s_{i+1}]}) = f_{\hat{y}}(\boldsymbol{x}_{(s_i,s_{i+1}]}) - \max_{y' \neq \hat{y}, y' \in \mathcal{Y}} f_{y'}(\boldsymbol{x}_{(s_i,s_{i+1}]}), \quad (5)$$

This importance score measures how far the prediction on a given feature is to the prediction boundary, hence the confidence of classifying $x_{(s_i,s_{i+1}]}$ into the predicted label $\hat{y}$.

Algorithm

Algorithm 1 Hierarchical Explanation via Divisive Generation

Input: text $\boldsymbol{x}$ with length n, and predicted label $\hat{y}$
Initialize the original partition $\mathcal{P}_0 \leftarrow \{\boldsymbol{x}_{(0,n]}\}$
Initialize the contribution set $\mathcal{C}_0 = \emptyset$
Initialize the hierarchy $\mathcal{H} = [\mathcal{P}_0]$
for $t = 1, \ldots, n-1$ **do**
    Find $\boldsymbol{x}_{(s_i,s_{i+1}]}$ and j by solving Equation 1
    Update the partition
    $\mathcal{P}'_t \leftarrow \mathcal{P}_{t-1} \backslash \{\boldsymbol{x}_{(s_i,s_{i+1}]}\}$
    $\mathcal{P}_t \leftarrow \mathcal{P}'_t \cup \{\boldsymbol{x}_{(s_i,j]}, \boldsymbol{x}_{(j,s_{i+1}]}\}$
    $\mathcal{H}.add(\mathcal{P}_t)$
    Update the contribution set $\mathcal{C}$ with
    $\mathcal{C}'_t \leftarrow \mathcal{C}_{t-1} \cup \{(\boldsymbol{x}_{(s_i,j]}, \psi(\boldsymbol{x}_{(s_i,j]}))\}$
    $\mathcal{C}_t \leftarrow \mathcal{C}'_t \cup \{(\boldsymbol{x}_{(j,s_{i+1}]}, \psi(\boldsymbol{x}_{(j,s_{i+1}]}))\}$
end for
Output: $\mathcal{C}_{n-1}, \mathcal{H}$

文本划分子集的贡献度集合

Experiments-实验数据与设置

The proposed method is evaluated on **text classification tasks** with three typical neural network models, **LSTM, CNN, and BERT**, **on the SST and IMDB datasets**, via both **automatic and human evaluations**.

Experiments- Quantitative Evaluation(定量分析)

1. AOPC: the area over the perturbation curve

2. log-odds scores

2. cohesion-score: to evaluate the interactions between words within a given text span.

Quantitative Evaluation -AOPC

$$\mathrm{AOPC}(k) = \frac{1}{N}\sum_{i=1}^{N}\{p(\hat{y} \mid \boldsymbol{x}_i) - p(\hat{y} \mid \tilde{\boldsymbol{x}}_i^{(k)})\},$$

Higher AOPCs are better, which means that the deleted words are important for model prediction.

Datasets	Methods	LSTM		CNN		BERT	
		AOPC	Log-odds	AOPC	Log-odds	AOPC	Log-odds
SST	Leave-one-out	0.441	-0.443	0.434	-0.448	0.464	-0.723
	CD	0.384	-0.382	-	-	-	-
	LIME	0.444	-0.449	0.473	-0.542	0.134	-0.186
	L-Shapley	0.431	-0.436	0.425	-0.459	0.435	-0.809
	C-Shapley	0.423	-0.425	0.415	-0.446	0.410	-0.754
	KernelSHAP	0.360	-0.361	0.387	-0.423	0.411	-0.765
	SampleShapley	0.450	-0.454	0.487	-0.550	0.462	-0.836
	HEDGE	**0.458**	**-0.466**	**0.494**	**-0.567**	**0.479**	**-0.862**
IMDB	Leave-one-out	0.630	-1.409	0.598	-0.806	0.335	-0.849
	CD	0.495	-1.190	-	-	-	-
	LIME	0.764	-1.810	0.691	-1.091	0.060	-0.133
	L-Shapley	0.637	-1.463	0.623	-0.950	0.347	-1.024
	C-Shapley	0.629	-1.427	0.613	-0.928	0.331	-0.973
	KernelSHAP	0.542	-1.261	0.464	-0.727	0.223	-0.917
	SampleShapley	0.757	-1.597	0.707	-1.108	0.355	-1.037
	HEDGE	**0.783**	**-1.873**	**0.719**	**-1.144**	**0.411**	**-1.126**

Quantitative Evaluation -Cohesion-score

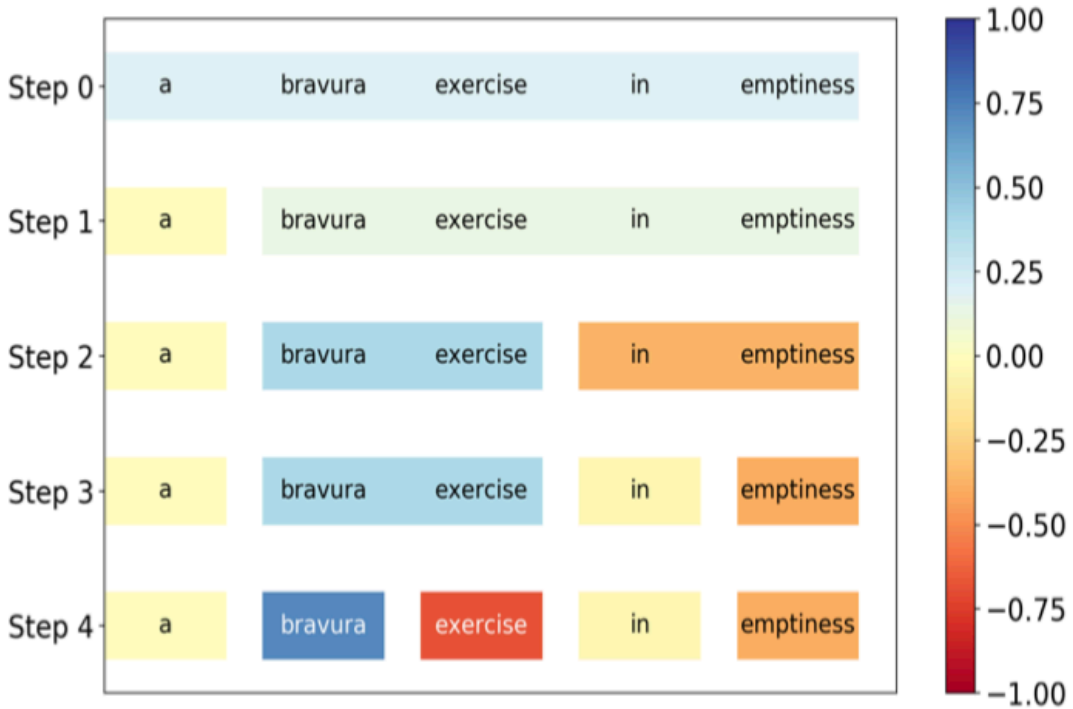
$$\text{Cohesion-score} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{Q}\sum_{q=1}^{Q}(p(\hat{y} \mid \boldsymbol{x}_i) - p(\hat{y} \mid \bar{\boldsymbol{x}}_i^{(q)}))$$

cohesion-score is the difference between $p(\hat{y}|x)$ and $p(\hat{y}|\bar{x})$.
通过对特征交互进行扰动，来观察预测概率下降的情况。

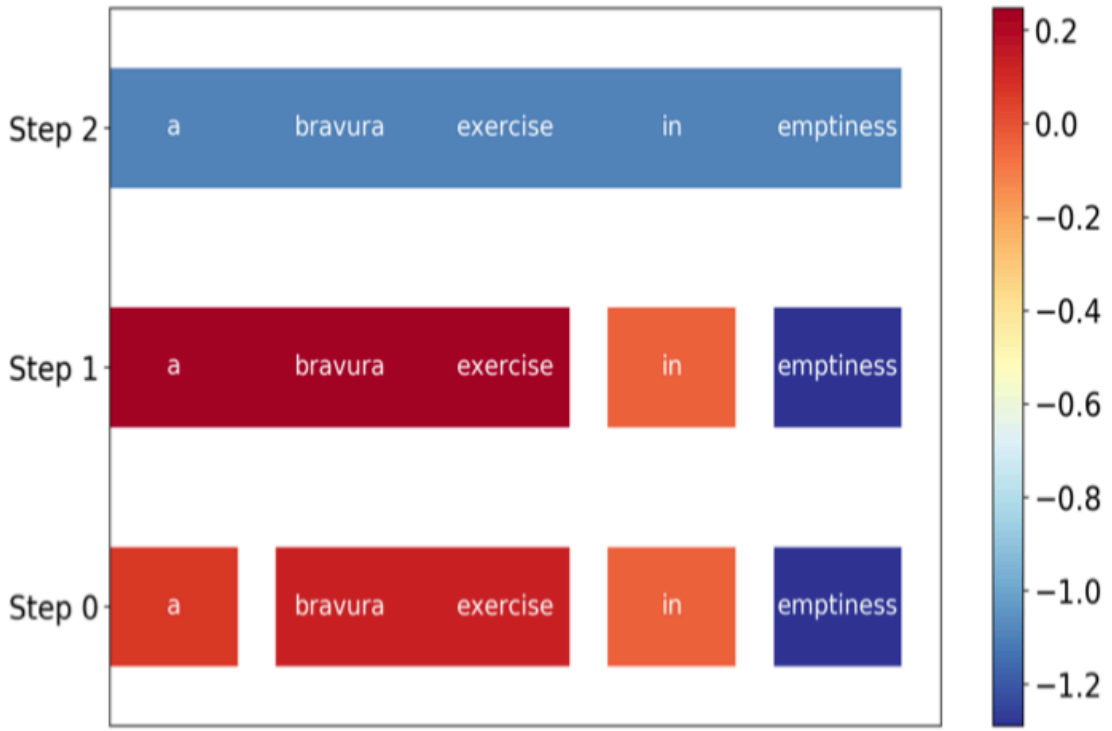
where $\bar{x}_i^{(q)}$ is the q th perturbed version of x_i , x_i 表示第i个样本

Methods	Models	Cohesion-score	
		SST	IMDB
HEDGE	CNN	0.016	0.012
	BERT	0.124	0.103
	LSTM	0.020	0.050
ACD	LSTM	0.015	0.038

Qualitative Analysis（定性分析） -Example 1



(a) HEDGE for LSTM on the SST.



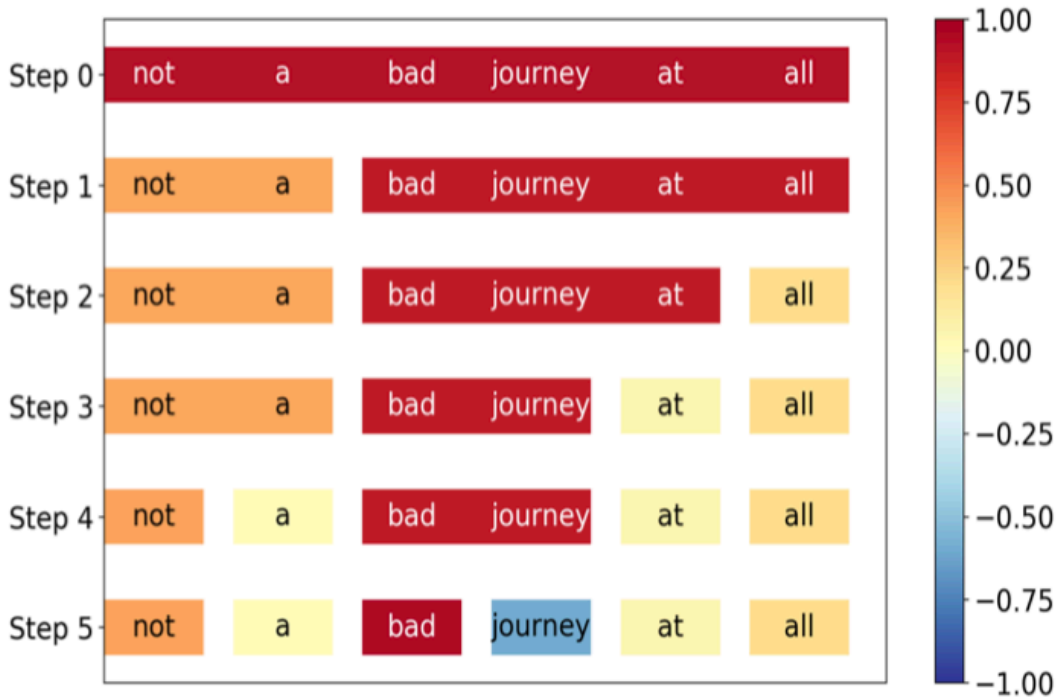
(b) ACD for LSTM on the SST.

HEDGE correctly captures the sentiment polarities of bravura and emptiness, and the interaction between them as bravura exercise flips the polarity of in emptiness to positive. It explains why the model makes the wrong prediction.

ACD incorrectly marks the two words with opposite polarities, and misses the feature interaction,

Qualitative Analysis（定性分析） -Example 2

Compare HEDGE in interpreting two different models (LSTM and BERT). BERT gives the correct prediction (POSITIVE), while LSTM makes a wrong prediction (NEGATIVE).



(a) HEDGE for LSTM on SST.

(b) HEDGE for BERT on SST.

LSTM misses the interaction between not and bad, and the negative word bad pushes the model making the NEGATIVE prediction.

BERT captures the key phrase not a bad at step 1, and thus makes the positive prediction,

Conclusion

(1) we design a top-down model-agnostic method of constructing hierarchical explanations via feature interaction detection;

(2) we propose a simple and effective scoring function to quantify feature contributions with respect to model predictions;

(3) we compare the proposed algorithm with several competitive methods on explanation generation via both automatic and human evaluations.

Thank U

Discussion

他不是真的不喜欢这个电影？