BlackBox NLP: What are we looking for, and where do we stand?

Sarah Wiegreffe
USC ISI NLP Seminar
January 30, 2020

Georgia Tech

Background

B.S. Data Science (2017)

Start PhD (2017)

Intern, Sutter Health (Summer 2018)

Intern, Google AI (Summer 2019)

present

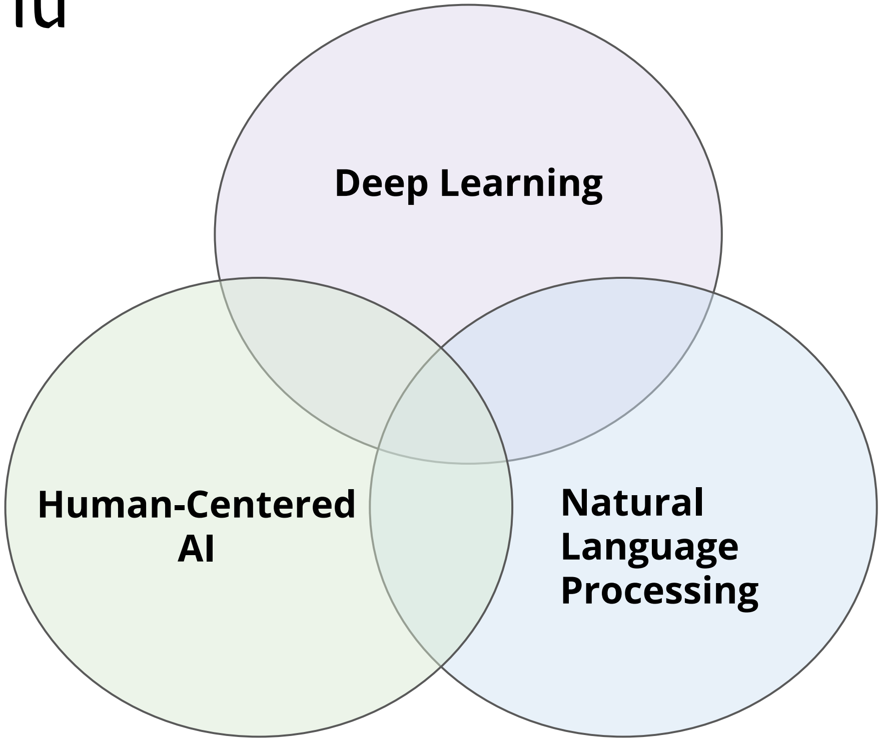
Intern, Allen AI (Summer 2020)

ML for Healthcare

Computational Linguistics

Explainable AI

Background


Deep Learning
Human-Centered AI
Natural Language Processing

Overview

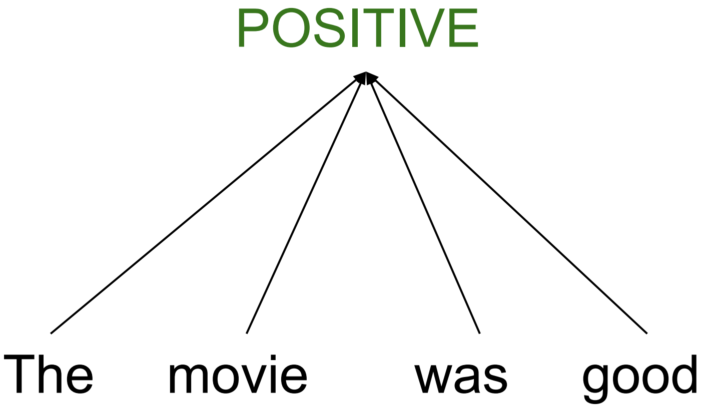
1. A Foray into Explainability
2. How do we define explanation?
3. Is attention explanation?
4. How do we guarantee faithfulness?
5. How do we test plausibility?
6. Future Directions

1. A Foray into Explainability
2. How do we define explanation?
3. Is attention explanation?
4. How do we guarantee faithfulness?
5. How do we test plausibility?
6. Future Directions

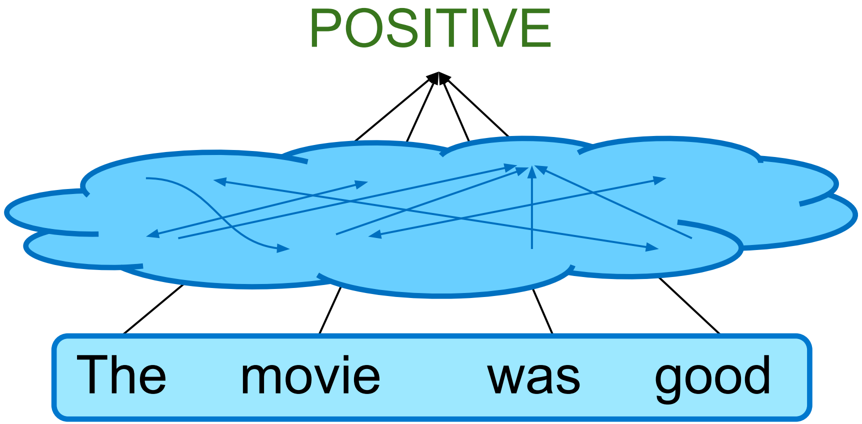
A Foray into Explainability

Mullenbach, Wiegreffe, Duke, Sun, Eisenstein. *Explainable Prediction of Clinical Codes from Clinical Text.* NAACL 2018.

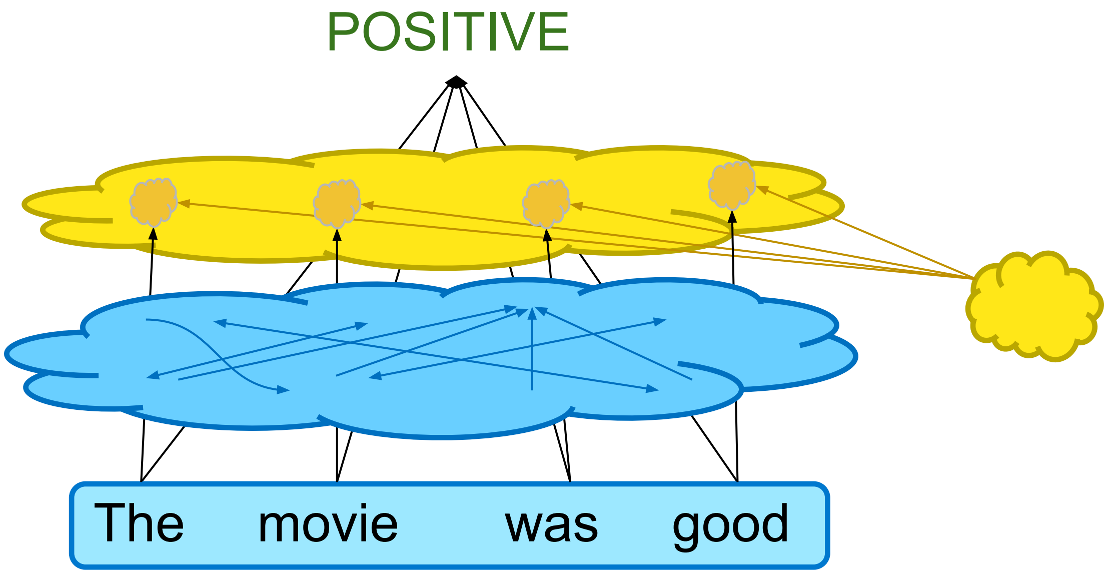
Classification Models



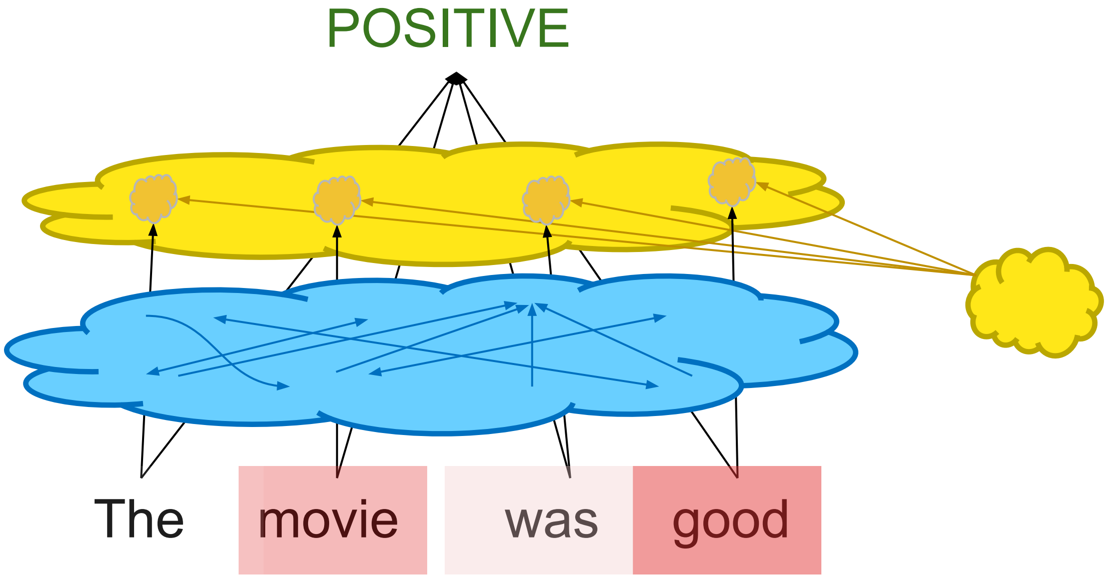
Neural Classification Models



Neural Classification Models with Attention



Neural Classification Models with Attention



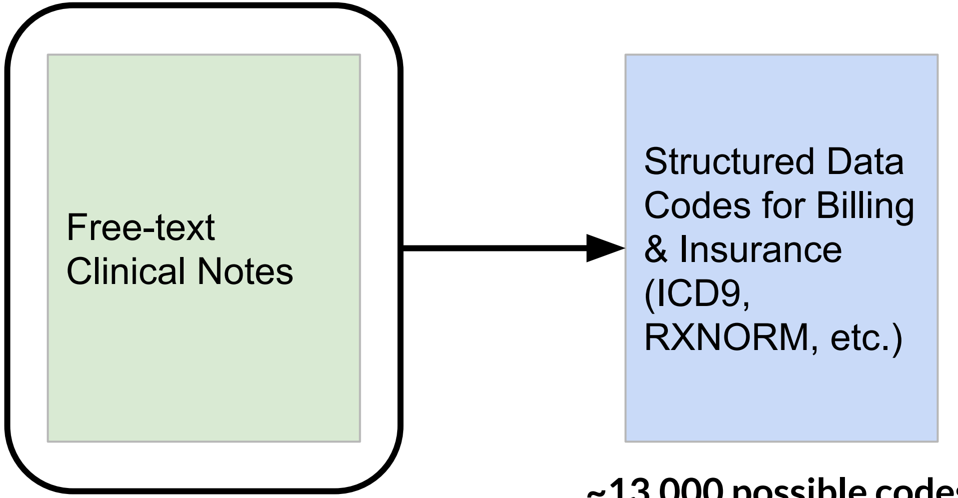
Clinical Coding Task

Free-text
Clinical Notes

Structured Data
Codes for Billing
& Insurance
(ICD9,
RXNORM, etc.)

Mullenbach, Wiegreffe, Duke, Sun, Eisenstein. *Explainable Prediction of Clinical Codes from Clinical Text.* NAACL 2018.

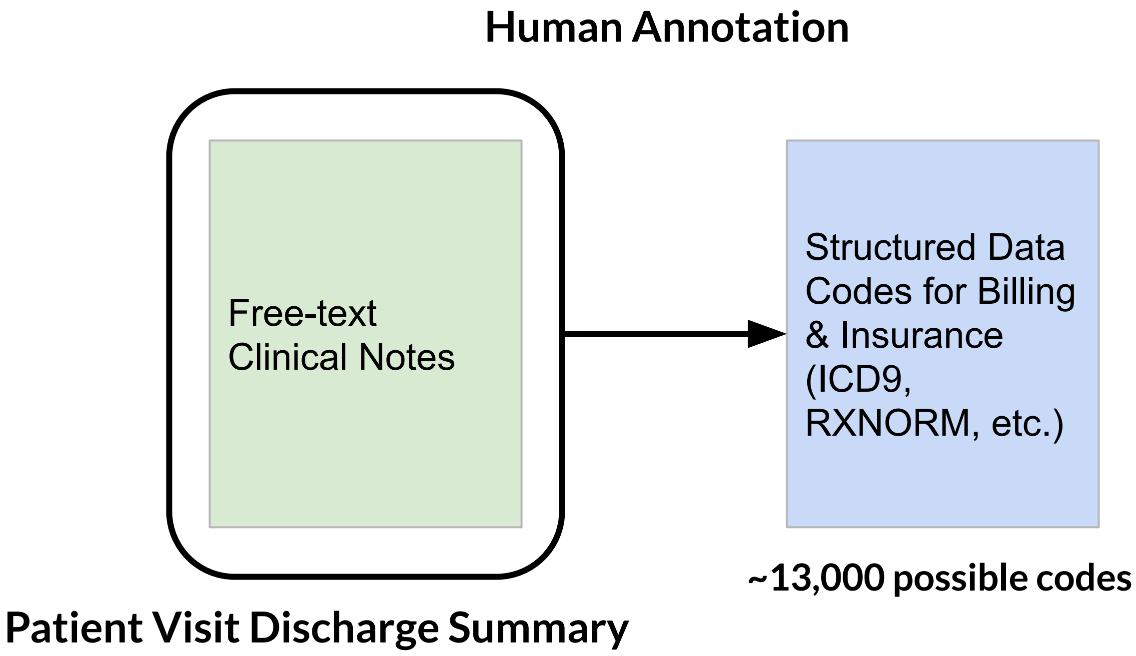
Clinical Coding Task



~13,000 possible codes

Patient Visit Discharge Summary

Clinical Coding Task



Clinical Coding Task

```
Admission Date:  [**2118-6-2**]       Discharge Date:  [**2118-6-14**]

Date of Birth:                        Sex:  F

Service:  MICU and then to [**Doctor Last Name **] Medicine

HISTORY OF PRESENT ILLNESS:  This is an 81-year-old female
with a history of emphysema (not on home O2), who presents...
```

```
519.1: 'Other disease…'
491.21: 'Obstructive …'
518.81: 'Acute respir…'
486: 'Pneumonia, orga…'
276.1: 'Hyposmolality…'
244.9: 'Unspecified h…'
31.99: 'Other operati…'
.
.
.
```

Motivation

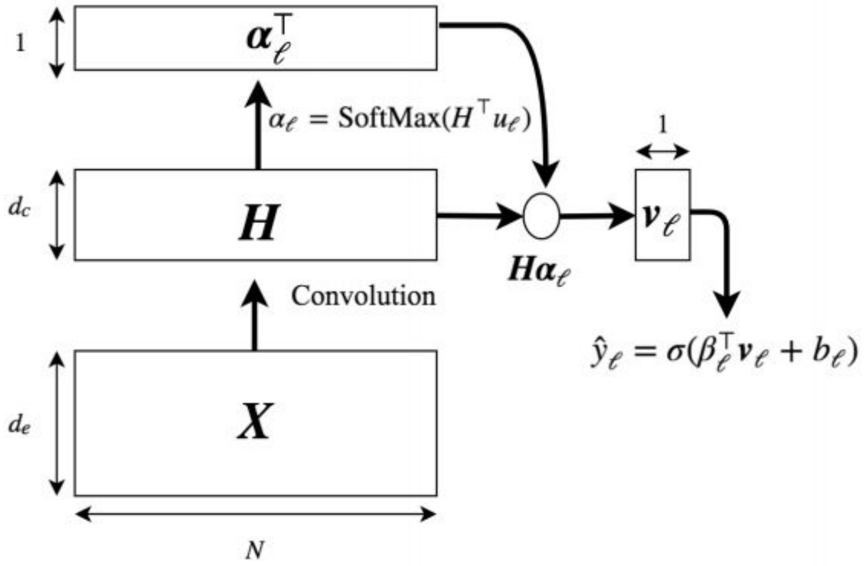
E849.0: Home accidents

801.26: ...subdural, and extradural hemorrhage...

...who sustained **a fall at home** she was found to have a large acute on **chronic subdural hematoma** with extensive midline shift...

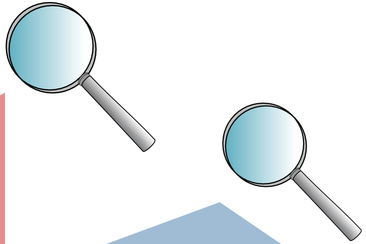
The CAML Model

- **C**onvolution + **A**ttention for **M**ulti-**L**abel classification
- Key Idea: *per-label* attention mechanism
- Achieved state-of-the-art on the ICD-9 clinical coding task



Physician Evaluation

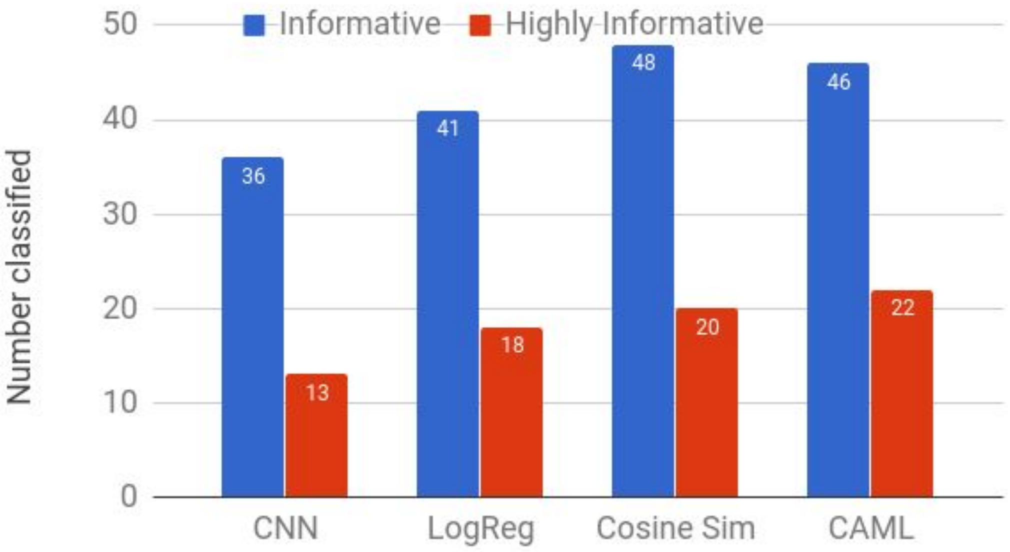
E849.0: Home accidents



801.26: ...subdural, and extradural hemorrhage...

...who sustained **a fall at home** she was found to have a large acute on **chronic subdural hematoma** with extensive midline shift...

Physician Evaluation



Were the extracted snippets *most responsible* for the model's prediction?

Were the extracted snippets *most responsible* for the model's prediction?

- Contextualization

Were the extracted snippets *most responsible* for the model's prediction?

- Contextualization
- Model variance

Were the extracted snippets *most responsible* for the model's prediction?

- Contextualization
- Model variance
- Exclusivity

Can Attention Weights provide explanation?

- An explanation is exclusive

Attention is not Explanation

Sarthak Jain **Byron C. Wallace**

Can Attention Weights provide explanation?

- An explanation is exclusive

Attention is not Explanation

Sarthak Jain **Byron C. Wallace**

- An explanation is robust

Is Attention Interpretable?

Sofia Serrano* **Noah A. Smith*†**

Can Attention Weights provide explanation?

- An explanation is exclusive

Attention is not Explanation

Sarthak Jain **Byron C. Wallace**

- An explanation is robust

Is Attention Interpretable?

Sofia Serrano* **Noah A. Smith*†**

The movie was good

Defining Explanation

Wiegreffe* and Pinter*. *Attention is not not Explanation.* EMNLP 2019.

Can Attention Weights provide explanation?

***Plausible* Explainability**

***Faithful* Explainability**

Can Attention Weights provide explanation?

Plausible Explainability

- Rationale generation
  (Ehsan et al. 2019, Riedl 2019)

Faithful Explainability

Can Attention Weights provide explanation?

***Plausible* Explainability**

- Rationale generation
  (Ehsan et al. 2019, Riedl 2019)

***Faithful* Explainability**

- Understanding correlation between inputs and output
  (Lipton 2016, Rudin 2018)

Can Attention Weights provide explanation?

Plausible Explainability

- Rationale generation
  (Ehsan et al. 2019, Riedl 2019)

Faithful Explainability

- Understanding correlation between inputs and output
  (Lipton 2016, Rudin 2018)
- Models' explanations are exclusive

Can Attention Weights provide explanation?

Plausible Explainability

- Rationale generation
  (Ehsan et al. 2019, Riedl 2019)

Faithful Explainability

- Understanding correlation between inputs and output
  (Lipton 2016, Rudin 2018)
- Models' explanations are exclusive

Can Attention Weights provide **faithful** explanation?

- A faithful explanation is exclusive

Attention is not Explanation

Sarthak Jain **Byron C. Wallace**

- A faithful explanation is robust

Is Attention Interpretable?

Sofia Serrano* **Noah A. Smith*†**

The movie was good

1. A Foray into Explainability
2. How do we define explanation?
3. Is attention explanation?
4. How do we guarantee faithfulness?
5. How do we test plausibility?
6. Future Directions

Is Attention (Faithful) Explanation?

Wiegreffe* and Pinter*. *Attention is not not Explanation.* EMNLP 2019.

If Attention is Faithful Explanation:

1. Attention should be a **necessary component** for good performance
2. If **trained models** can vary in attention distributions while giving similar predictions, they might be bad for explanation
3. Attention weights should work well in **uncontextualized settings**

Necessary

Hard to manipulate

Work out of context

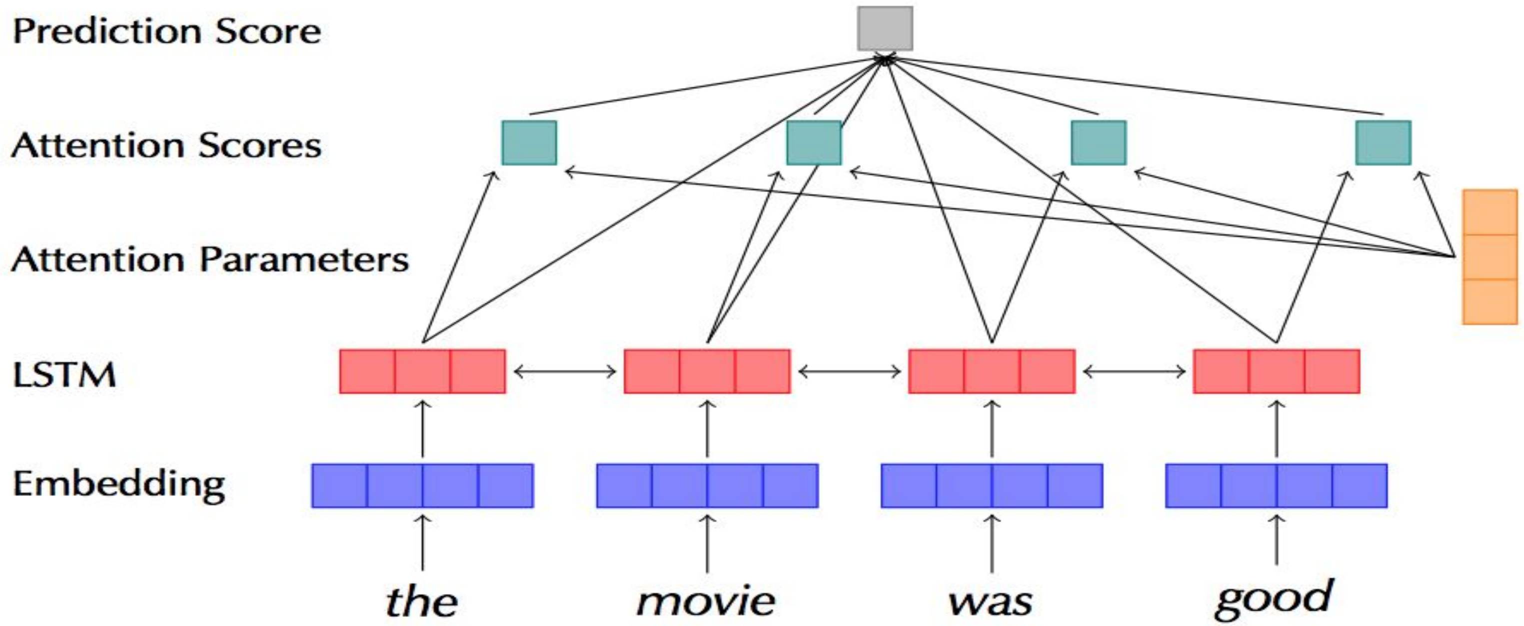
Selecting Meaningful Tasks

Necessary

1. Attention should be a **necessary component** for good performance

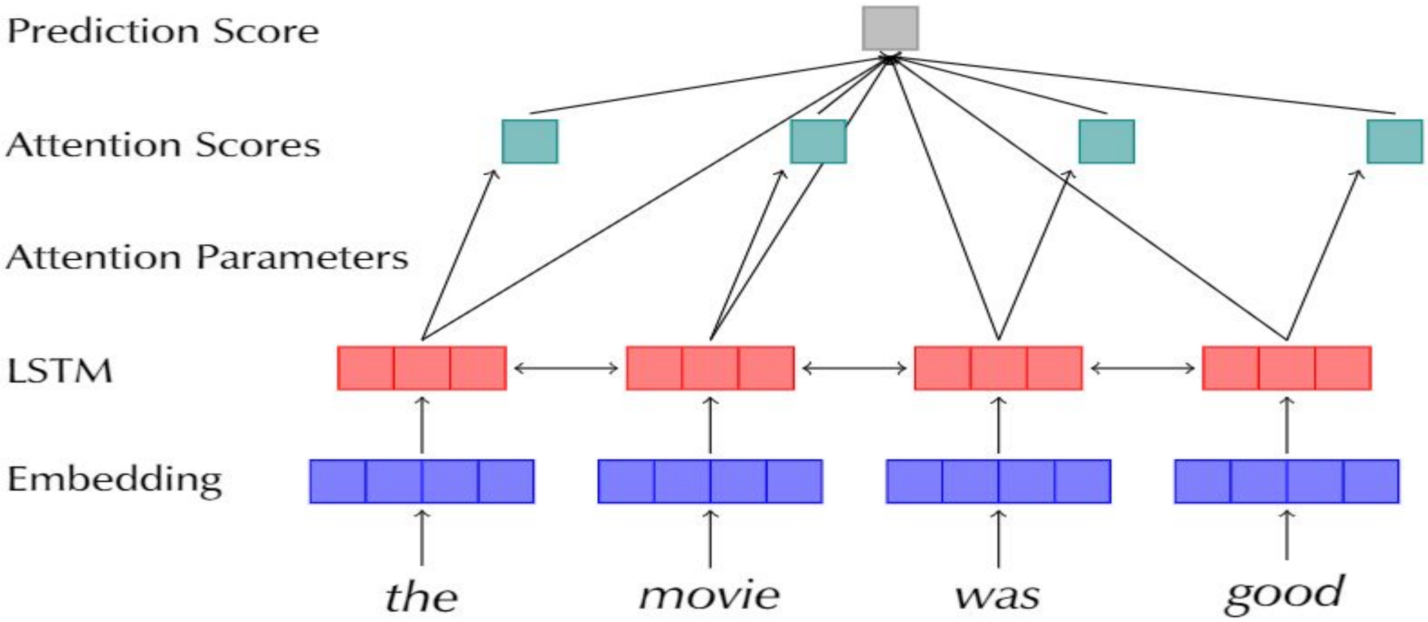
Selecting Meaningful Tasks

Necessary



Selecting Meaningful Tasks

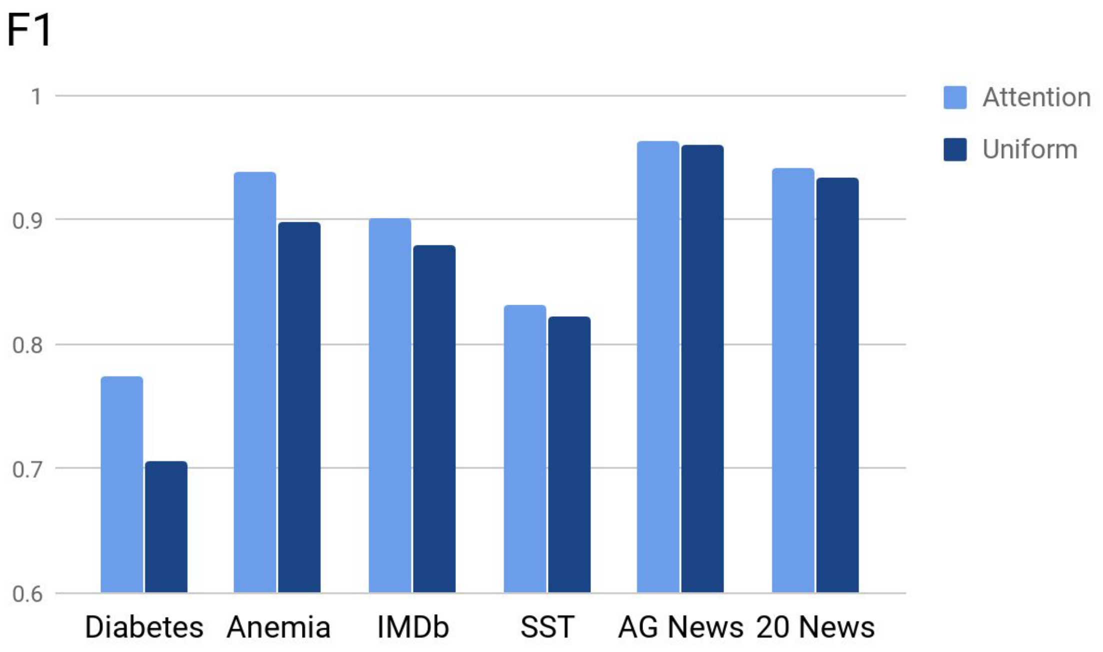
Necessary



Set to Uniform

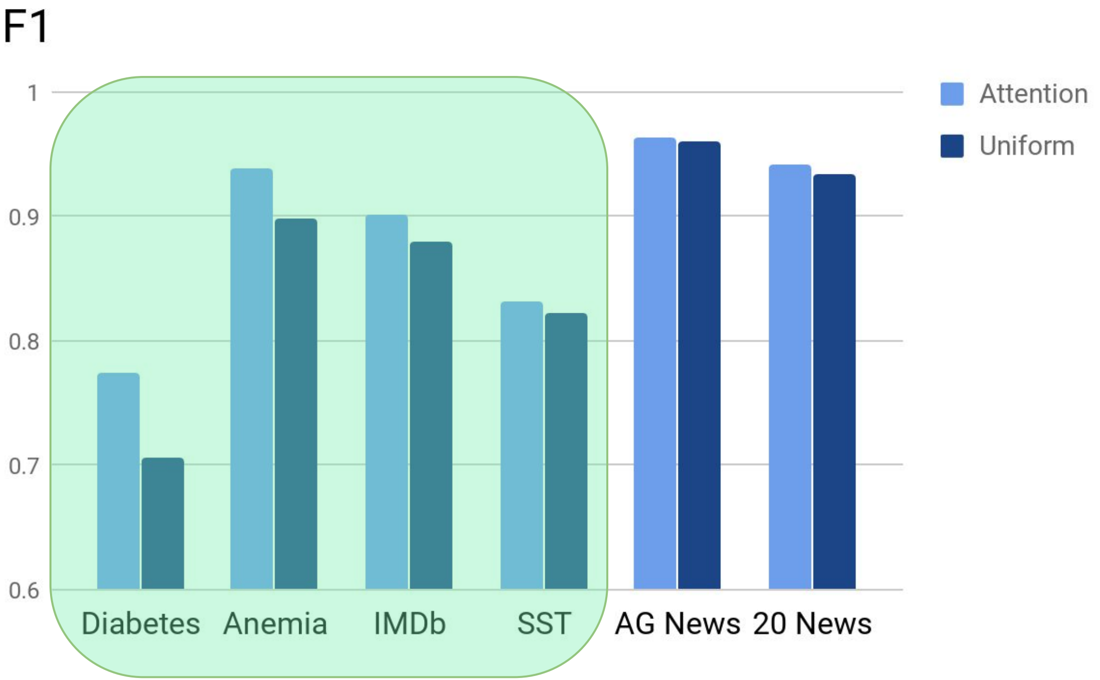
Selecting Meaningful Tasks

Necessary



Selecting Meaningful Tasks

Necessary



Searching for Adversarial Models

Hard to manipulate

1. Attention should be a **necessary component** for good performance

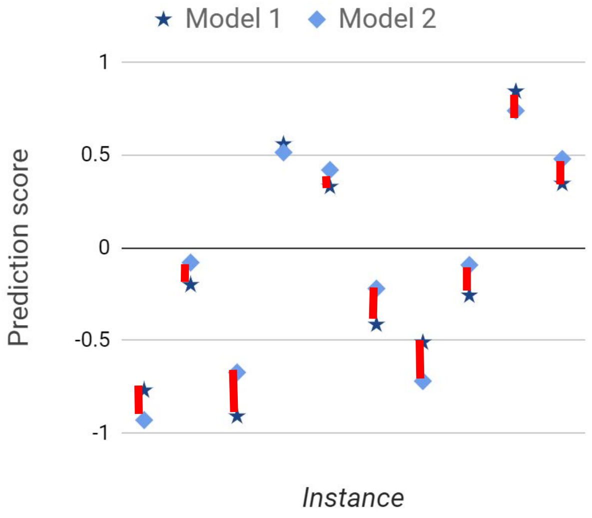
2. If **trained models** can vary in attention distributions while giving similar predictions, they might be bad for explanation

Measures

Hard to manipulate

- Total Variation Distance: for comparing class predictions between 2 models

$$\text{TVD}(\hat{y}_1, \hat{y}_2) = \frac{1}{2}\sum_{i=1}^{|\mathcal{Y}|} |\hat{y}_{1i} - \hat{y}_{2i}|$$



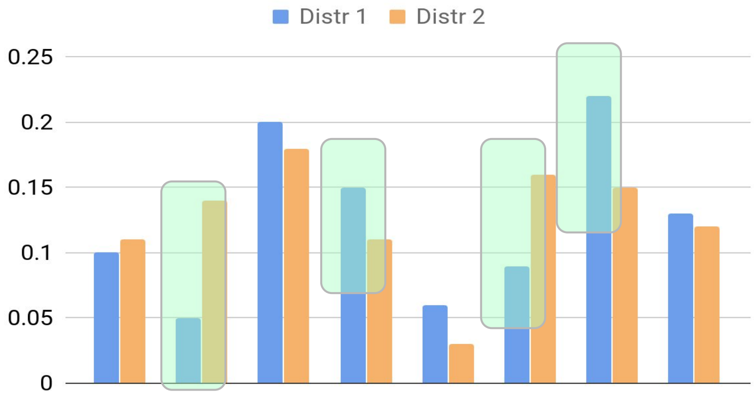
Measures

Hard to manipulate

- Jensen–Shannon Divergence: for comparing 2 distributions

$$\mathrm{JSD}(\alpha_1, \alpha_2) = \frac{1}{2}\,\mathrm{KL}[\alpha_1 \parallel \bar{\alpha}] + \frac{1}{2}\,\mathrm{KL}[\alpha_2 \parallel \bar{\alpha}],$$

where $\bar{\alpha} = \frac{\alpha_1 + \alpha_2}{2}$.



Adversarial Training

Hard to manipulate

1. Train a base model (M_b)
2. Train an adversary (M_a) that **minimizes change in prediction scores** from the base model, while *maximizing changes in the learned attention distributions.*

$$\mathcal{L}(\mathcal{M}_a, \mathcal{M}_b)^{(i)} = \mathrm{TVD}(\hat{y}_a^{(i)}, \hat{y}_b^{(i)}) - \lambda \ \mathrm{KL}(\boldsymbol{\alpha}_a^{(\mathbf{i})} \parallel \boldsymbol{\alpha}_b^{(\mathbf{i})})$$

Adversarial Training

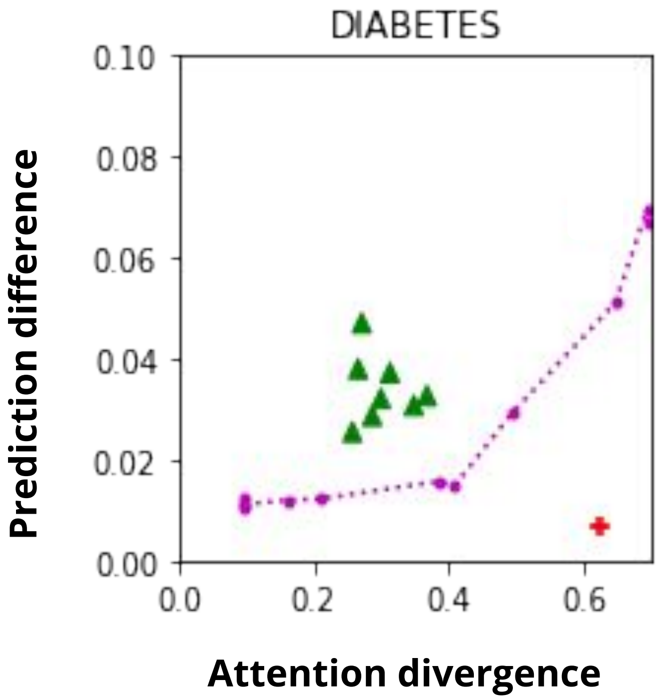
Hard to manipulate

1. Train a base model (M_b)
2. Train an adversary (M_a) that **minimizes change in prediction scores** from the base model, while *maximizing changes in the learned attention distributions.*

$$\mathcal{L}(\mathcal{M}_a, \mathcal{M}_b)^{(i)} = \mathrm{TVD}(\hat{y}_a^{(i)}, \hat{y}_b^{(i)}) - \boxed{\lambda}\, \mathrm{KL}(\boldsymbol{\alpha}_a^{(\mathbf{i})} \parallel \boldsymbol{\alpha}_b^{(\mathbf{i})})$$

Adversarial Results

Hard to manipulate

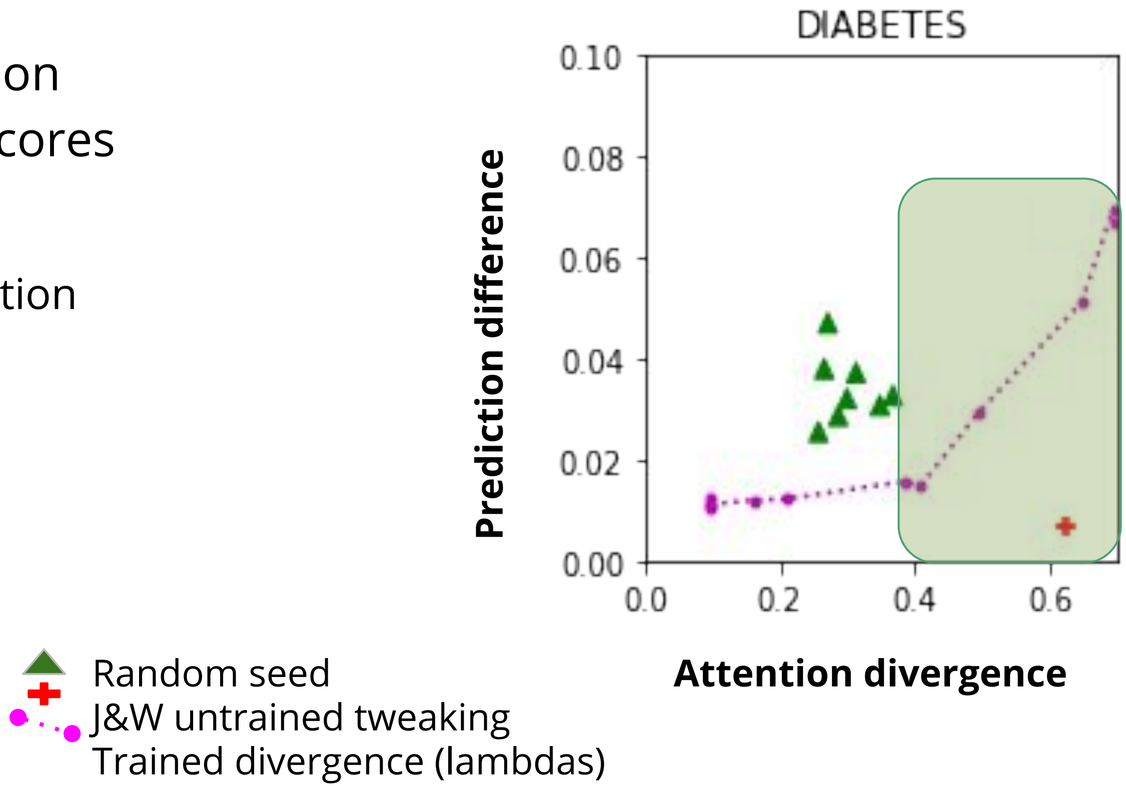


Random seed
J&W untrained tweaking
Trained divergence (lambdas)

Adversarial Results

Hard to manipulate

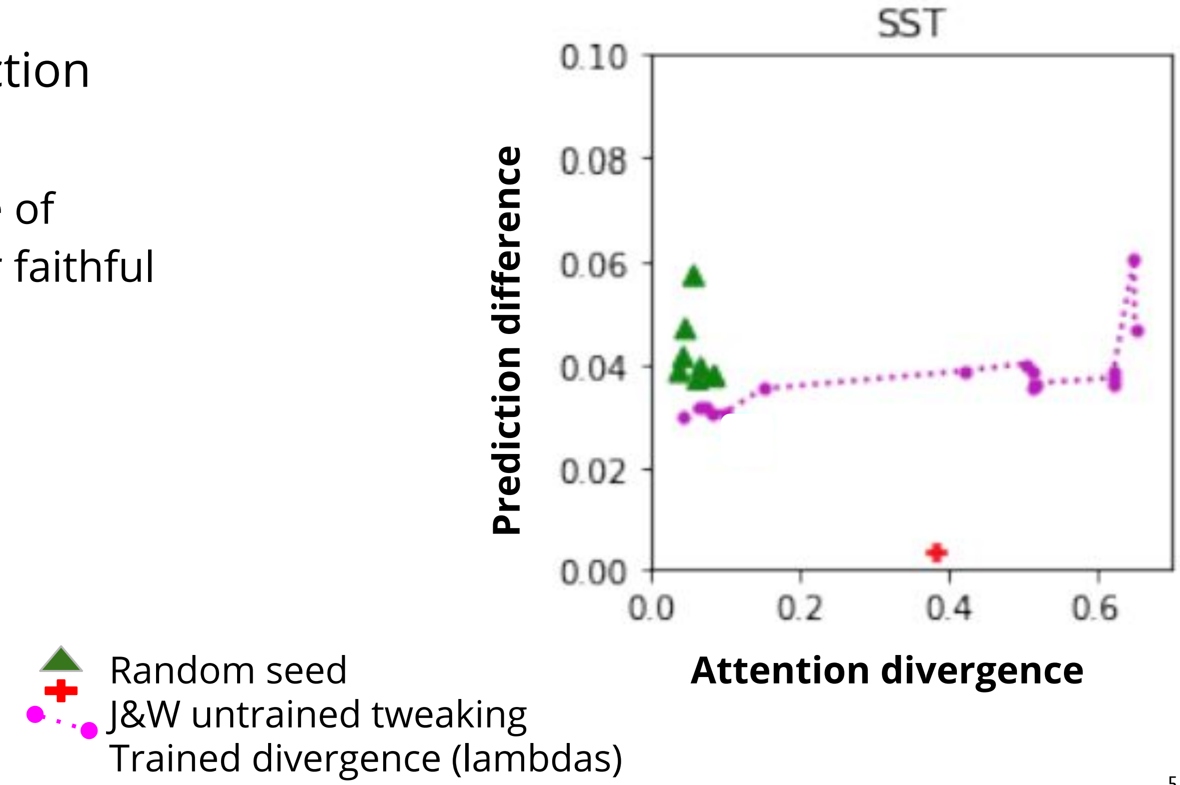
- Fast increase in prediction difference = attention scores not easily manipulable
  - Supports use of attention weights for faithful explanation



Adversarial Results

Hard to manipulate

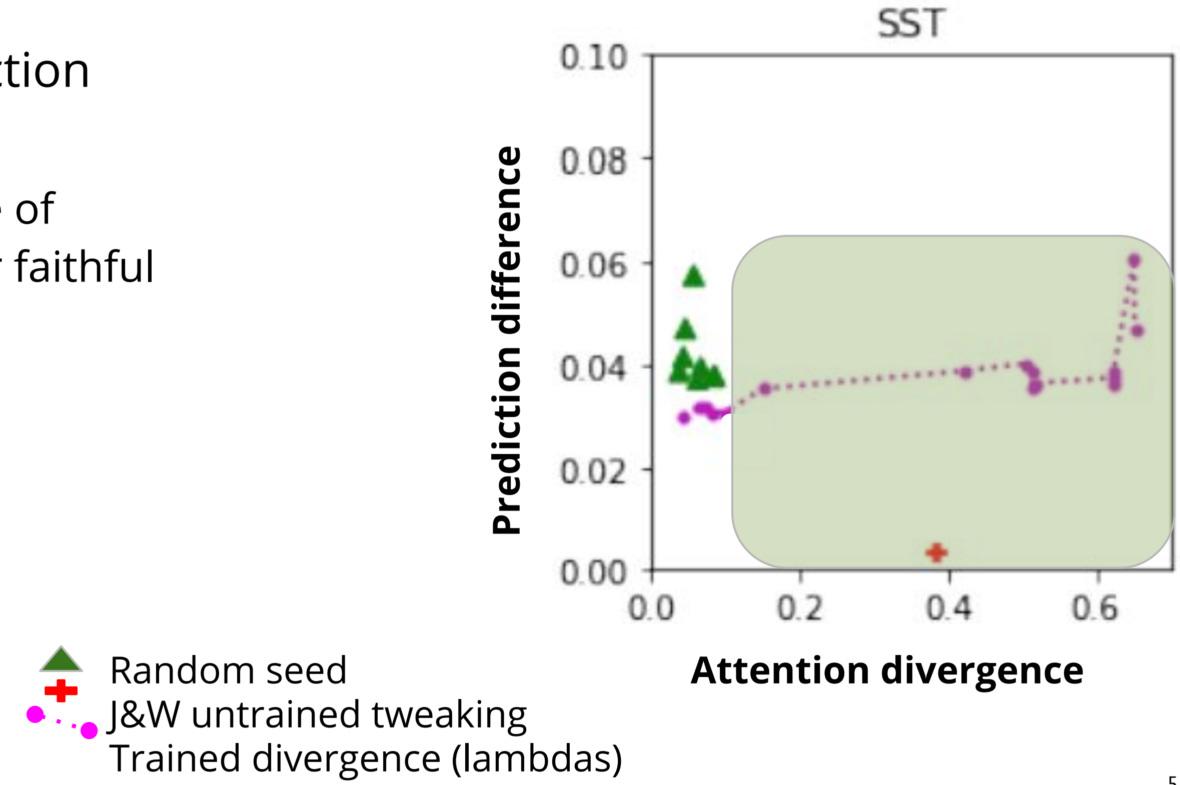
- Slow increase in prediction difference
    - *Does not* support use of attention weights for faithful explanation



Adversarial Results

Hard to manipulate

- Slow increase in prediction difference
    - *Does not* support use of attention weights for faithful explanation



Probing Attention

Work out of context

1. Attention should be a **necessary component** for good performance

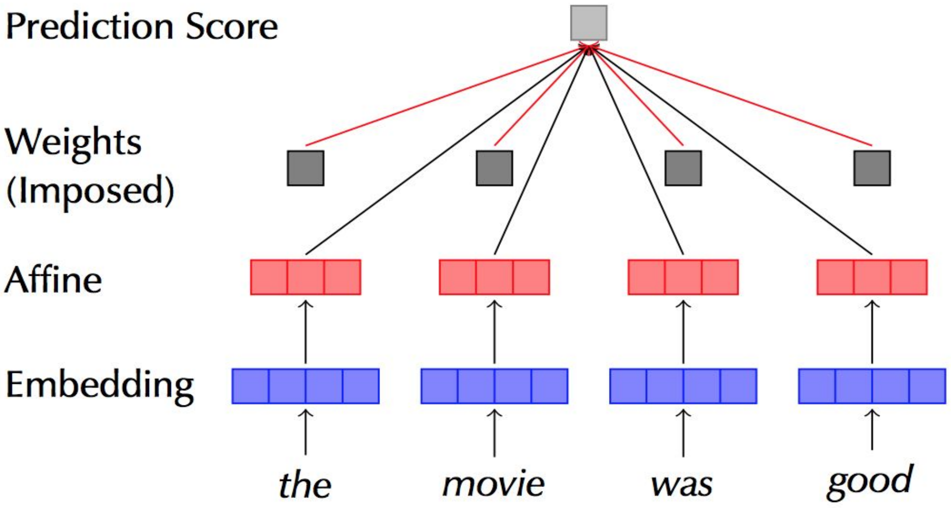
2. If **trained models** can vary in attention distributions while giving similar predictions, they might be bad for explanation

3. Attention weights should work well in **uncontextualized settings**

Probing Attention

Work out of context

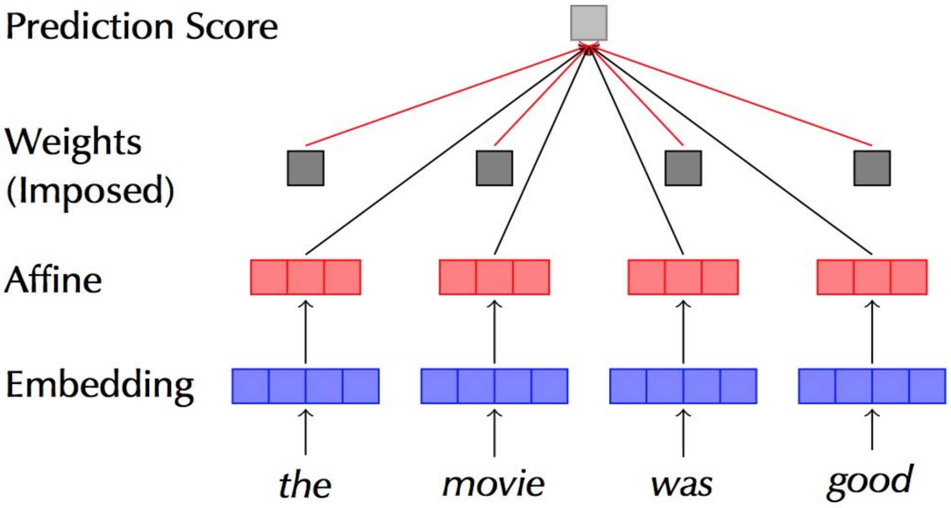
- Treat the learned attention weights as a **guide** in a non-contextualized, bag-of-word-vectors model



Probing Attention

Work out of context

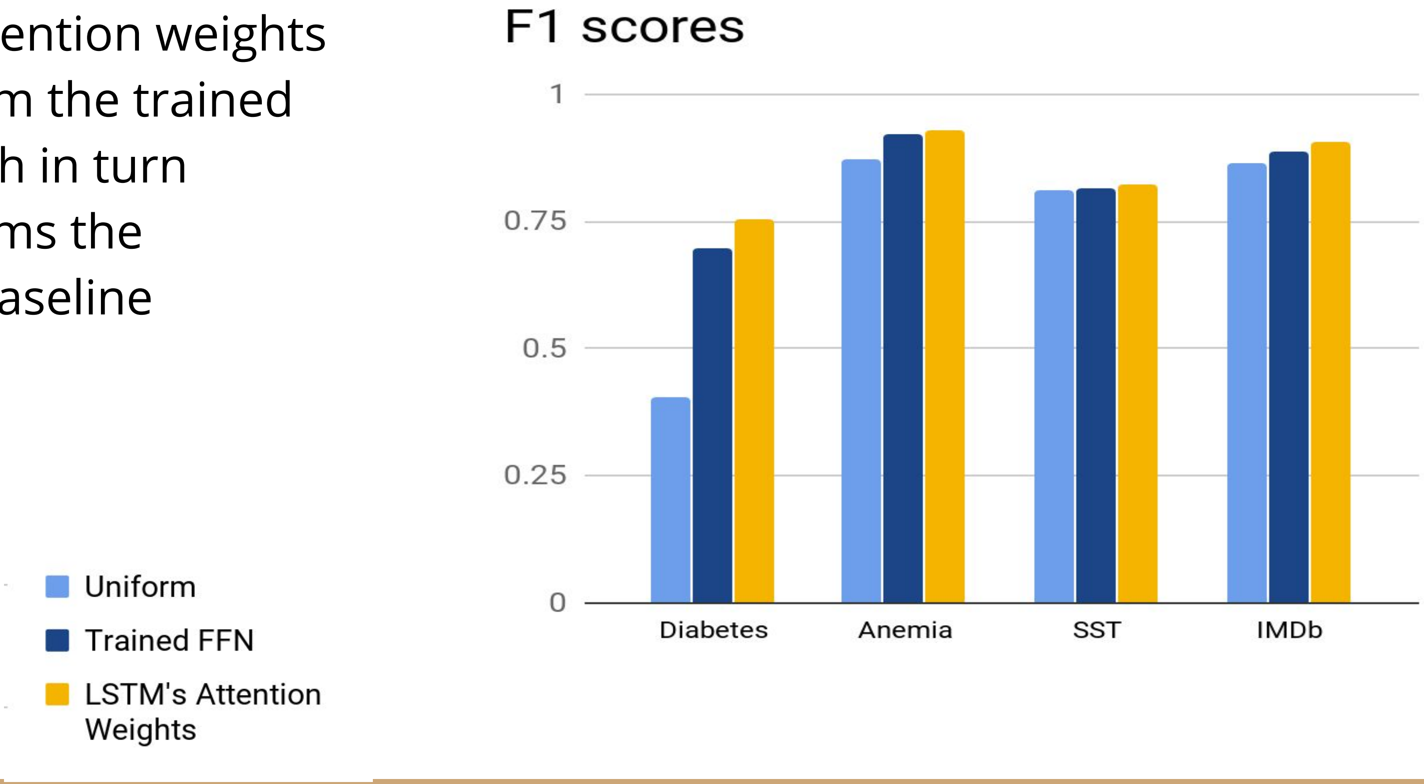
- Treat the learned attention weights as a **guide** in a non-contextualized, bag-of-word-vectors model
- High performance → attention scores capture relationship between inputs and output



Results

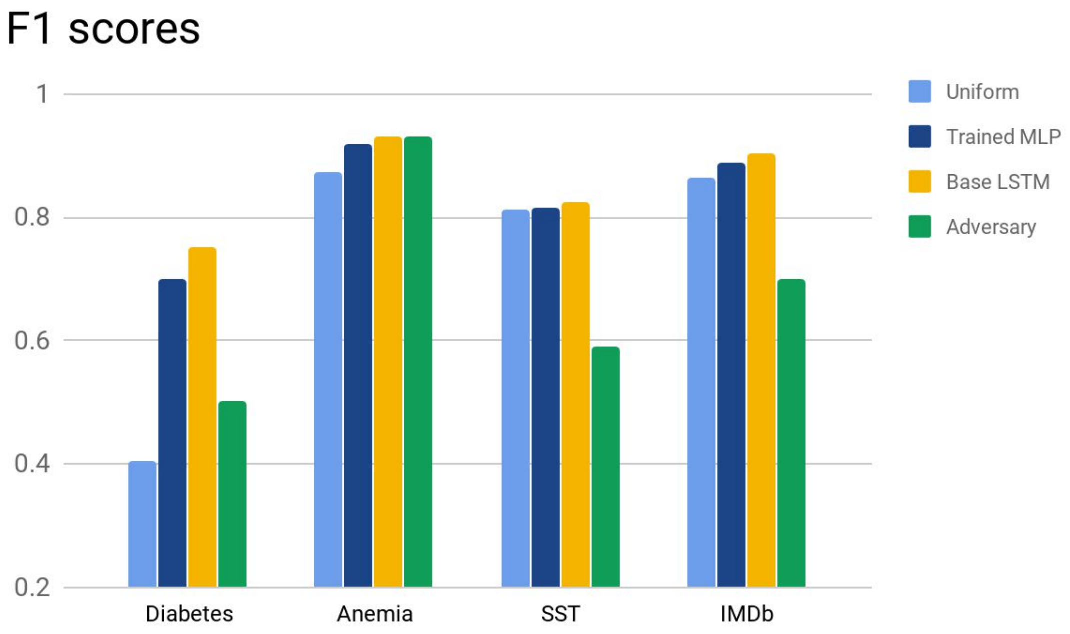
Work out of context

- LSTM's attention weights outperform the trained MLP, which in turn outperforms the uniform baseline



Results

Work out of context



Conclusion- Is Attention Explanation?

- 3 desiderata of attention for “faithful” explanation

Necessary

Hard to manipulate

Work out of context

Conclusion- is Attention Explanation?

- 3 desiderata of attention for “faithful” explanation
- 3 methods to measure the utility of attention distributions for faithful explanation

Necessary	Select Meaningful Tasks
Hard to manipulate	Search for Adversaries
Work out of context	Use Attention as Guides

Conclusion- Is Attention Explanation?

- 3 desiderata of attention for “faithful” explanation
- 3 methods to measure the utility of attention distributions for faithful explanation
- Results showing performance is highly task-dependent

Necessary	Select Meaningful Tasks
Hard to manipulate	Search for Adversaries
Work out of context	Use Attention as Guides

Guaranteeing Faithfulness by Construction

Jain, Wiegreffe, Pinter, Wallace. Under review.

Background

- Explanation as an (extractive) subset-selection problem

Background

- Explanation as an (extractive) subset-selection problem
- Lei et al.* propose to jointly train rationale generation and task prediction modules

*Tao Lei, Regina Barzilay, and Tommi Jaakkola. *Rationalizing Neural Predictions.* EMNLP 2016.

Background

- Explanation as an (extractive) subset-selection problem
- Lei et al.* propose to jointly train rationale generation and task prediction modules
    - Discrete nature of method necessitates training via REINFORCE
    - High variance, necessitates careful hyperparameter tuning
    - Difficult to adopt in practice

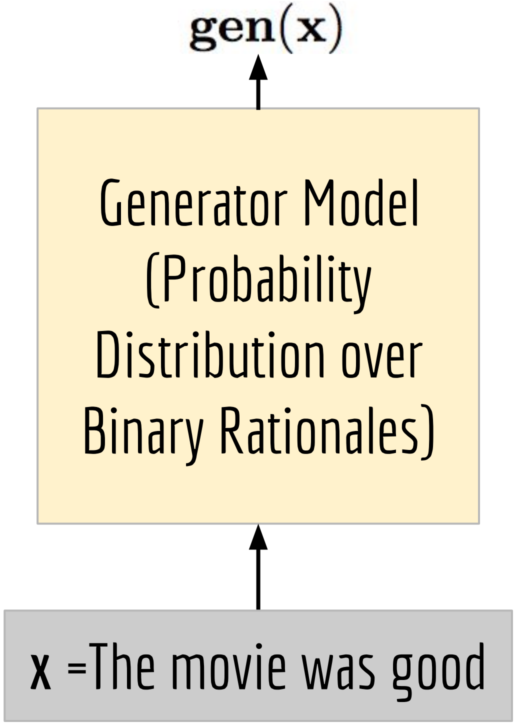
*Tao Lei, Regina Barzilay, and Tommi Jaakkola. *Rationalizing Neural Predictions.* EMNLP 2016.

Background

- Explanation as an (extractive) subset-selection problem
- Lei et al.* propose to jointly train rationale generation and task prediction modules
    - Discrete nature of method necessitates training via REINFORCE
    - High variance, necessitates careful hyperparameter tuning
    - Difficult to adopt in practice
- Constrained prediction *guarantees faithfulness*

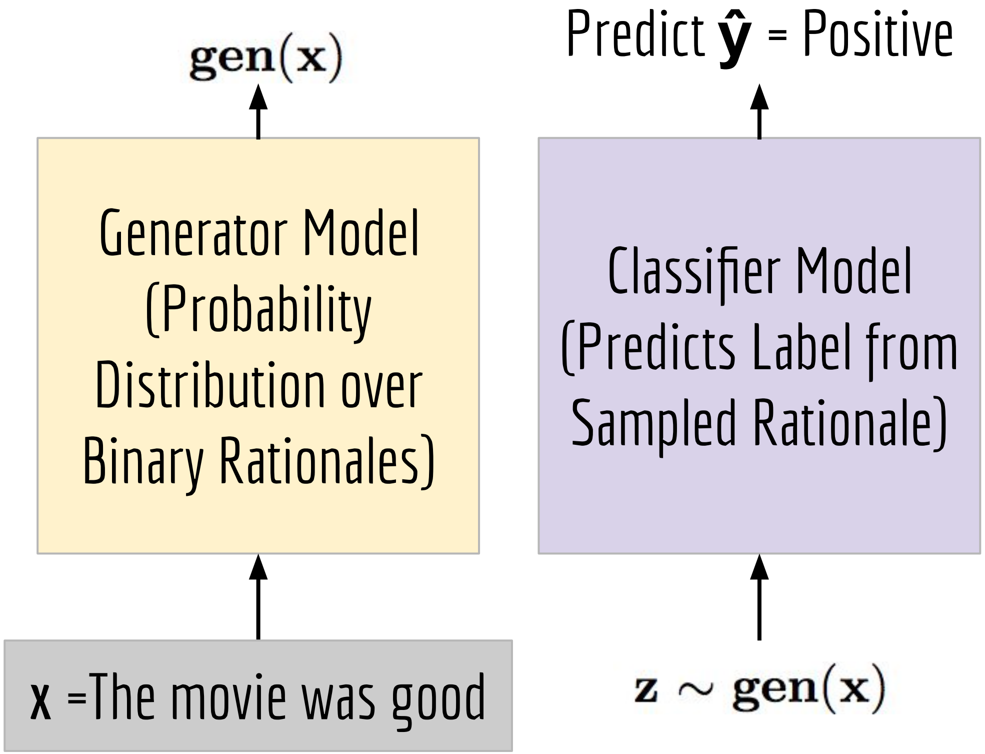
*Tao Lei, Regina Barzilay, and Tommi Jaakkola. *Rationalizing Neural Predictions.* EMNLP 2016.

Lei et al. Model



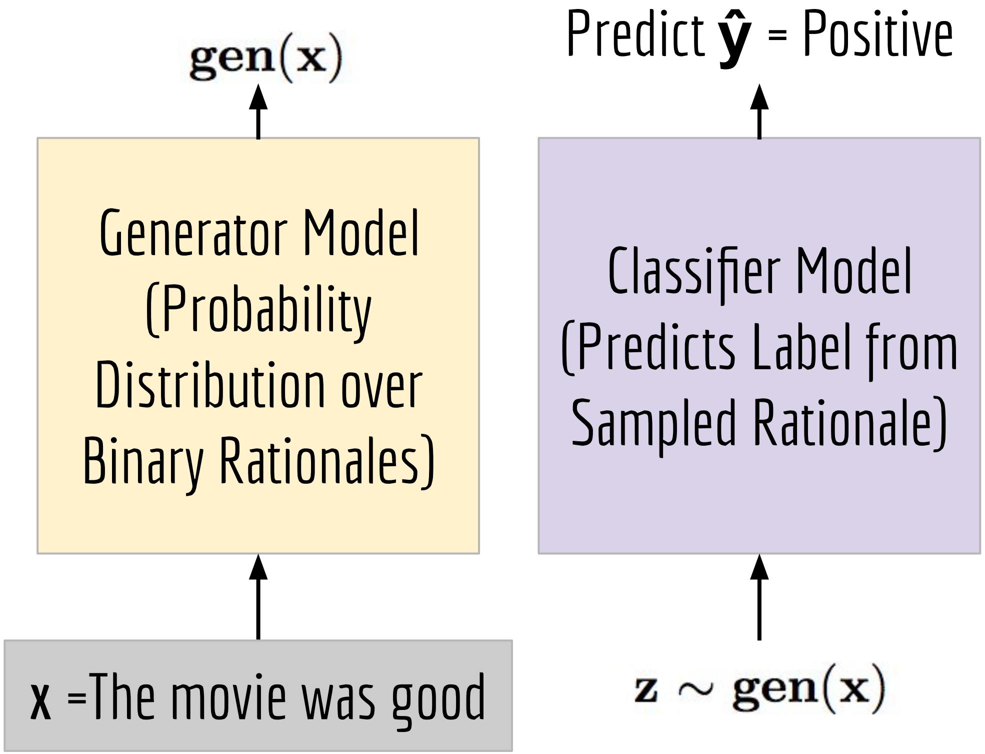
Tao Lei, Regina Barzilay, and Tommi Jaakkola. *Rationalizing Neural Predictions.* EMNLP 2016.

Lei et al. Model



Tao Lei, Regina Barzilay, and Tommi Jaakkola. *Rationalizing Neural Predictions.* EMNLP 2016.

Lei et al. Model



$$\min_{\theta_e, \theta_g} \sum_{(\mathbf{x},\mathbf{y}) \in D} \mathbb{E}_{\mathbf{z} \sim \mathbf{gen}(\mathbf{x})} \left[\mathrm{cost}(\mathbf{z}, \mathbf{x}, \mathbf{y}) \right]$$

Tao Lei, Regina Barzilay, and Tommi Jaakkola. *Rationalizing Neural Predictions.* EMNLP 2016.

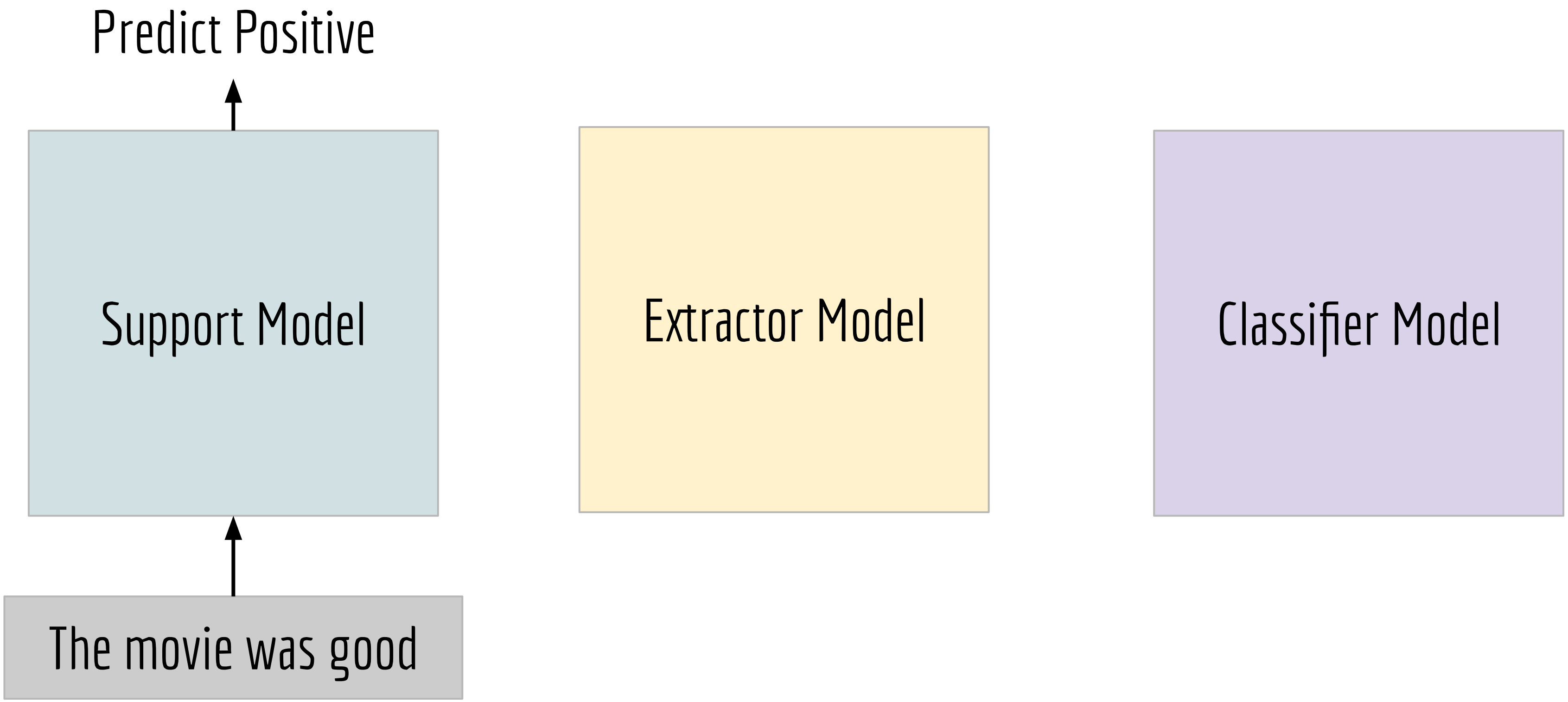
FRESH Model

Support Model

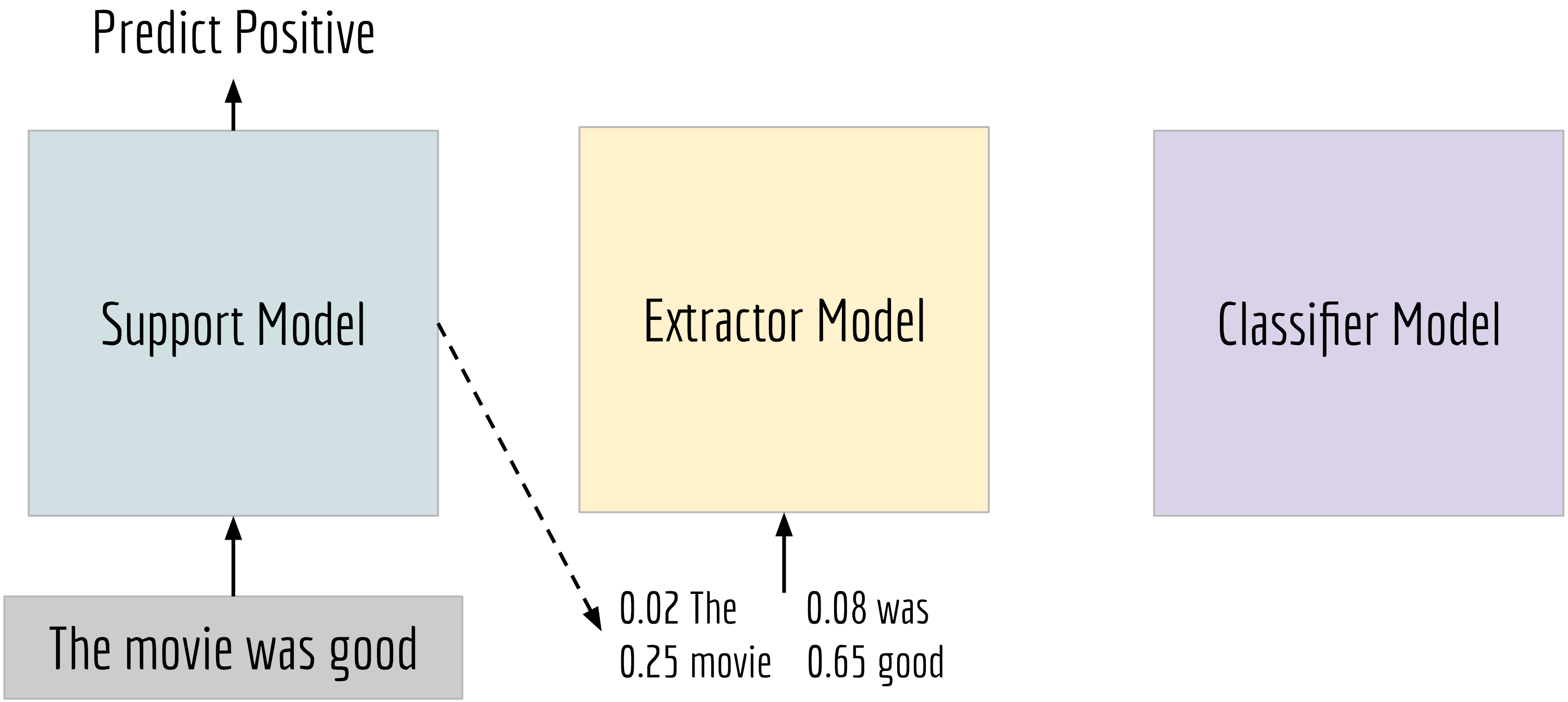
Extractor Model

Classifier Model

FRESH Model

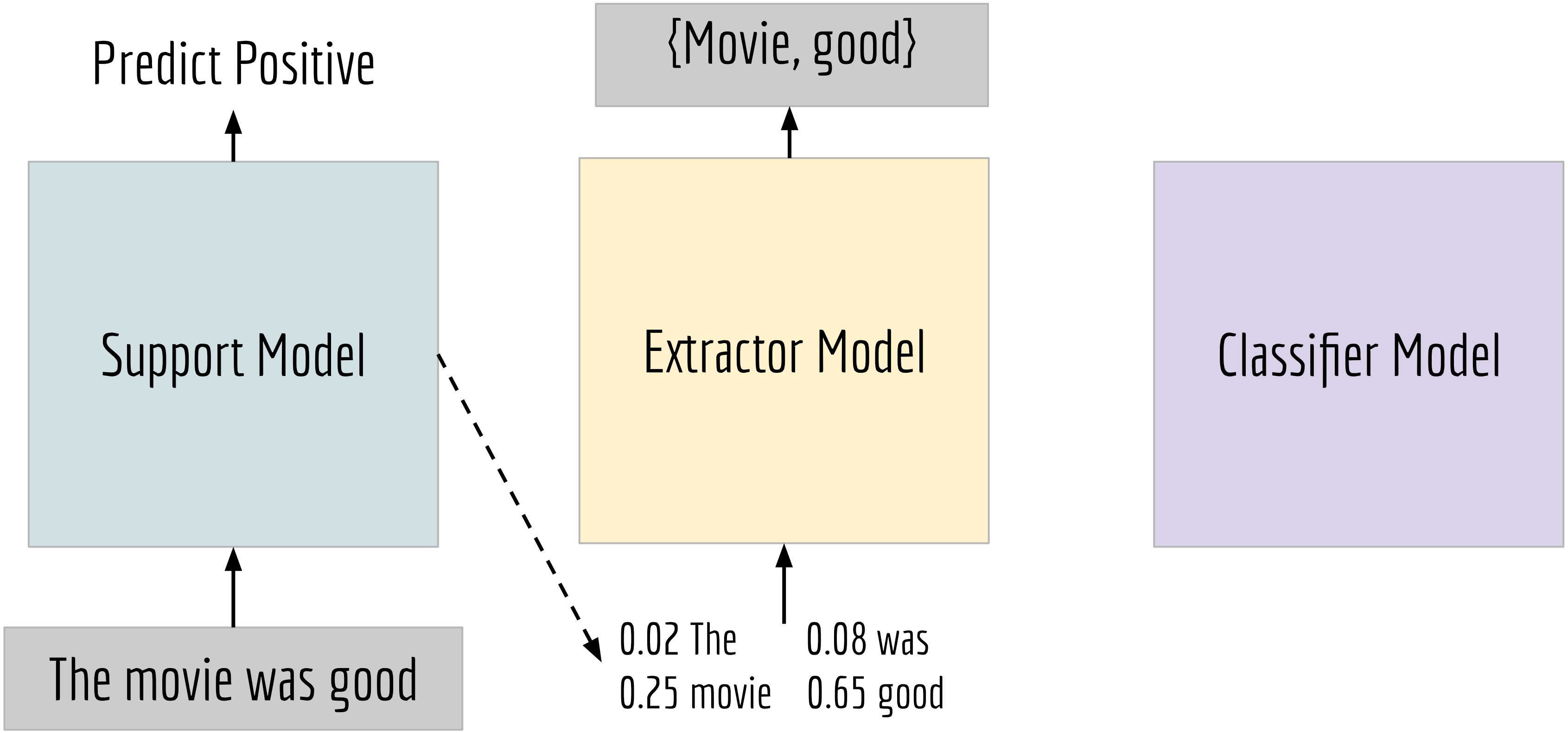

Predict Positive
Support Model
Extractor Model
Classifier Model
The movie was good

FRESH Model

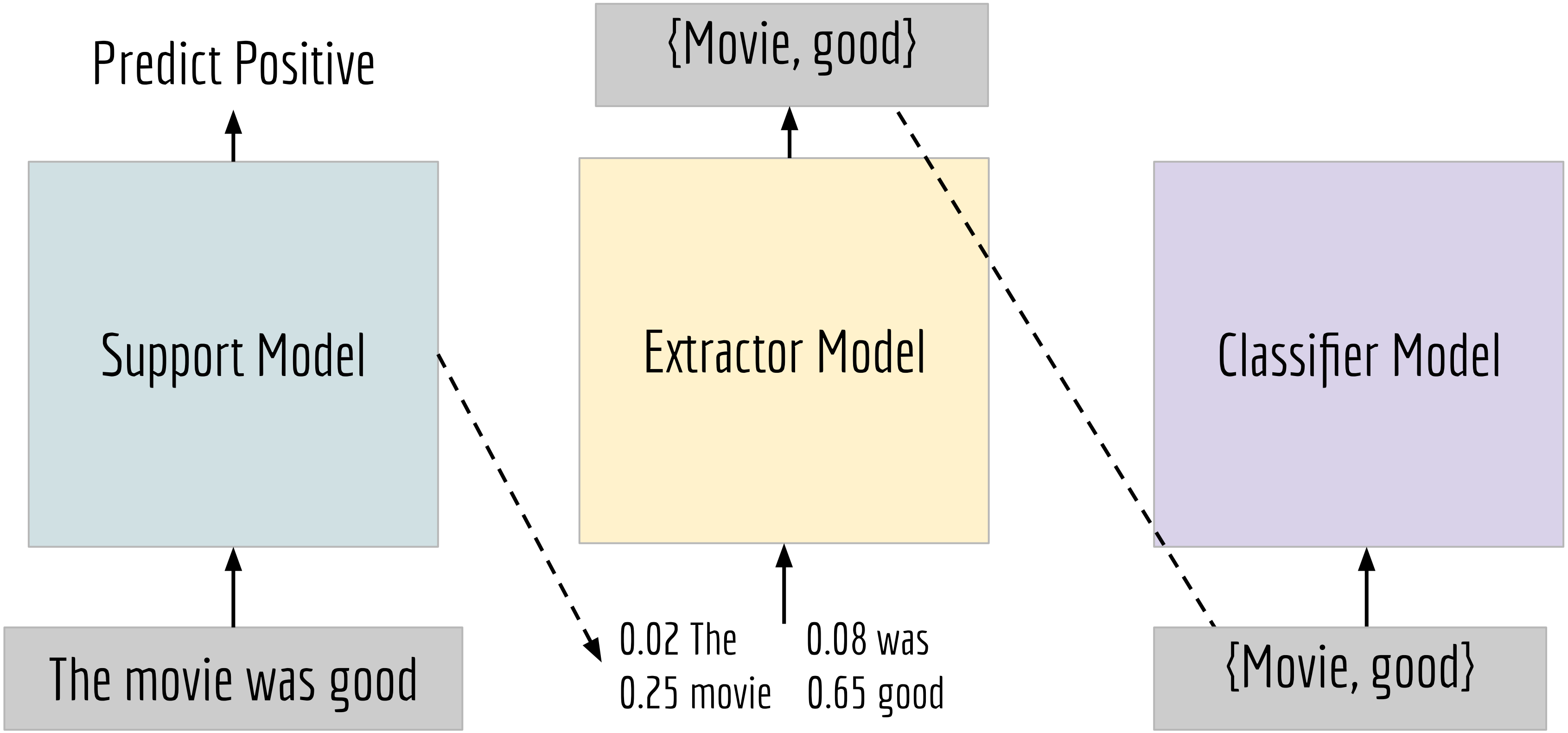


Predict Positive
Support Model
Extractor Model
Classifier Model
The movie was good
0.02 The
0.25 movie
0.08 was
0.65 good

FRESH Model

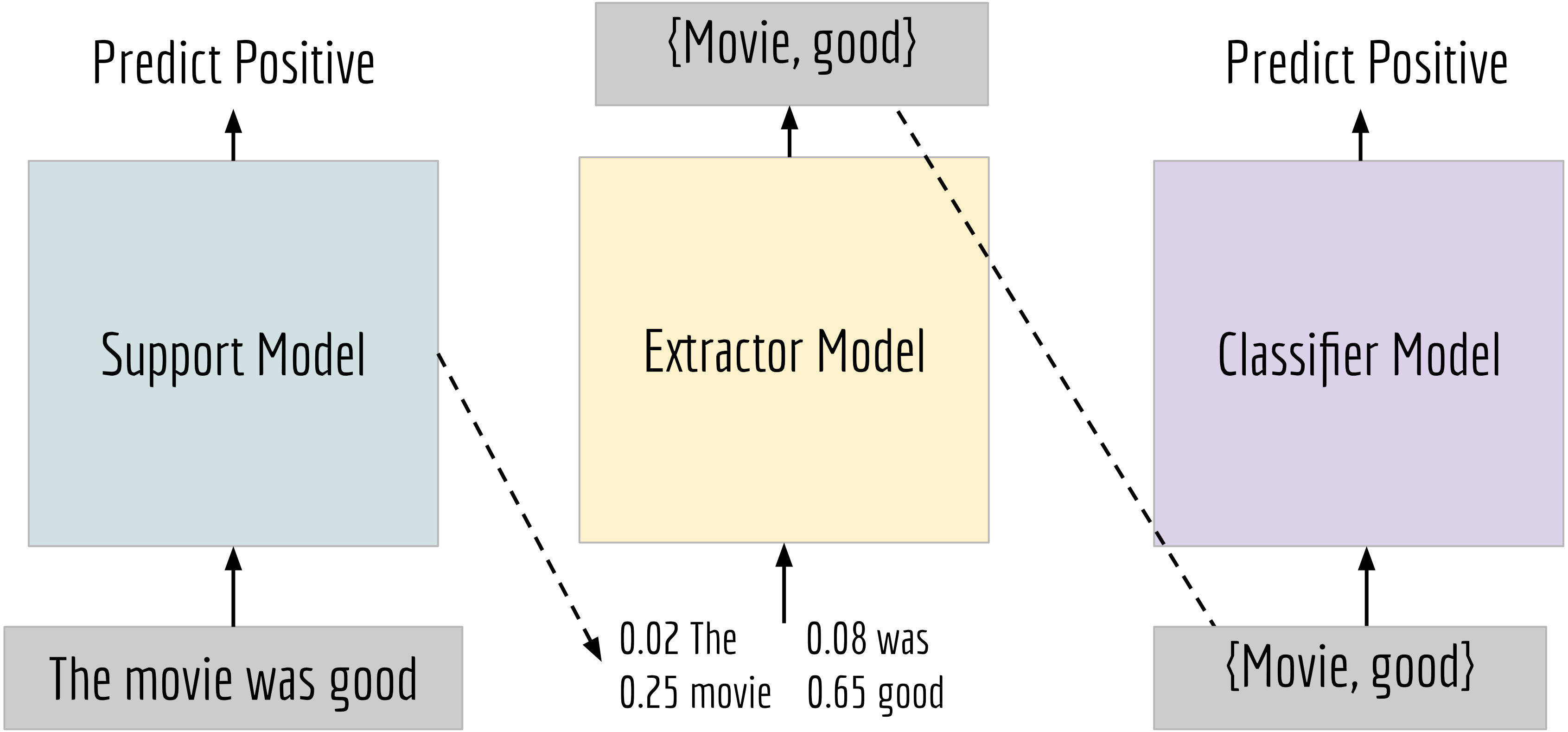

Predict Positive
{Movie, good}
Support Model
Extractor Model
Classifier Model
The movie was good
0.02 The
0.25 movie
0.08 was
0.65 good

FRESH Model



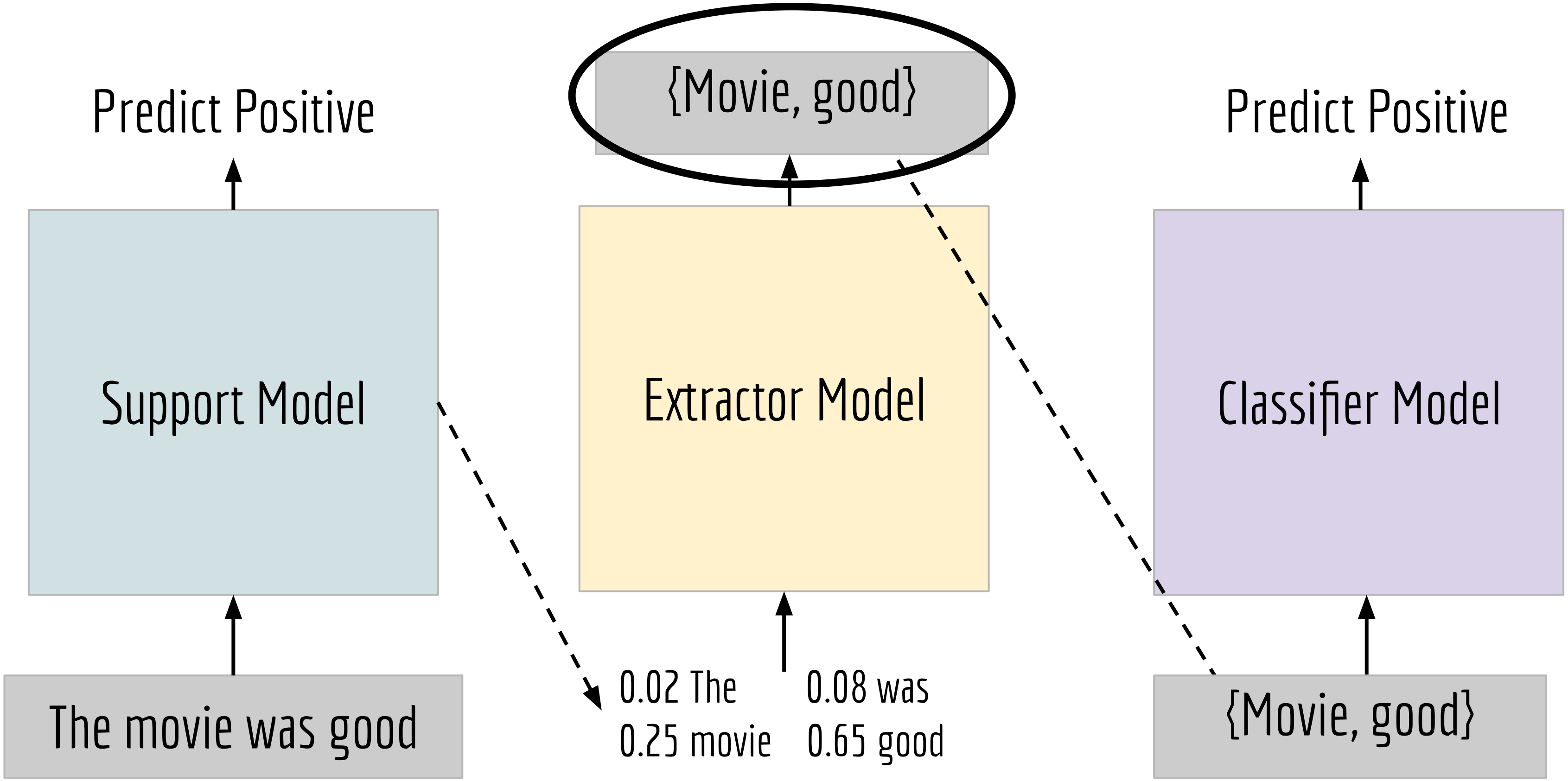
Predict Positive
{Movie, good}
Support Model
Extractor Model
Classifier Model
The movie was good
0.02 The
0.08 was
0.25 movie
0.65 good
{Movie, good}

FRESH Model



Predict Positive
Support Model
The movie was good
{Movie, good}
Extractor Model
0.02 The
0.25 movie
0.08 was
0.65 good
Predict Positive
Classifier Model
{Movie, good}

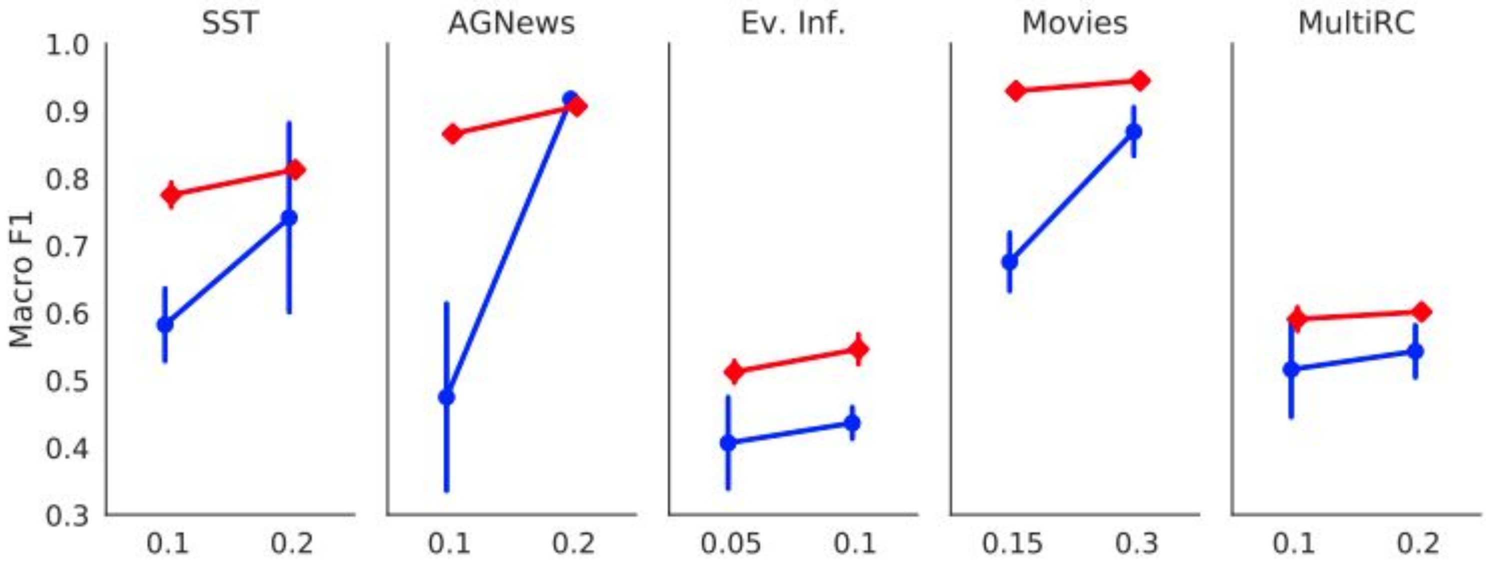
FRESH Model


{Movie, good}
Predict Positive
Predict Positive
Support Model
Extractor Model
Classifier Model
The movie was good
0.02 The
0.25 movie
0.08 was
0.65 good
{Movie, good}

Results

- FRESH outperforms prior models, recovering most of the performance of the original black box.
- FRESH achieves better average performance than the end-to-end method.

Results



Proportion of Document used for rationale

Testing Plausibility with Human Evaluations

Jain, Wiegreffe, Pinter, Wallace. Under review.

Human Evaluations

Plausible Explainability

- Rationale generation
  (Ehsan et al. 2019, Riedl 2019)

Faithful Explainability

- Understanding correlation between inputs and output
  (Lipton 2016, Rudin 2018)
- Models' explanations are exclusive

Human Evaluations

Plausible Explainability

- Rationale generation
  (Ehsan et al. 2019, Riedl 2019)

Faithful Explainability

- Understanding correlation between inputs and output
  (Lipton 2016, Rudin 2018)
- Models' explanations are exclusive

Rationale Plausibility

Sufficiency

- Can a human predict the correct label given only the rationale? (Kim et al. 2016)
- In our model: can a human perform the task of the Classifier module?

Rationale Plausibility

Sufficiency

- Can a human predict the correct label given only the rationale? (Kim et al. 2016)
- In our model: can a human perform the task of the Classifier module?

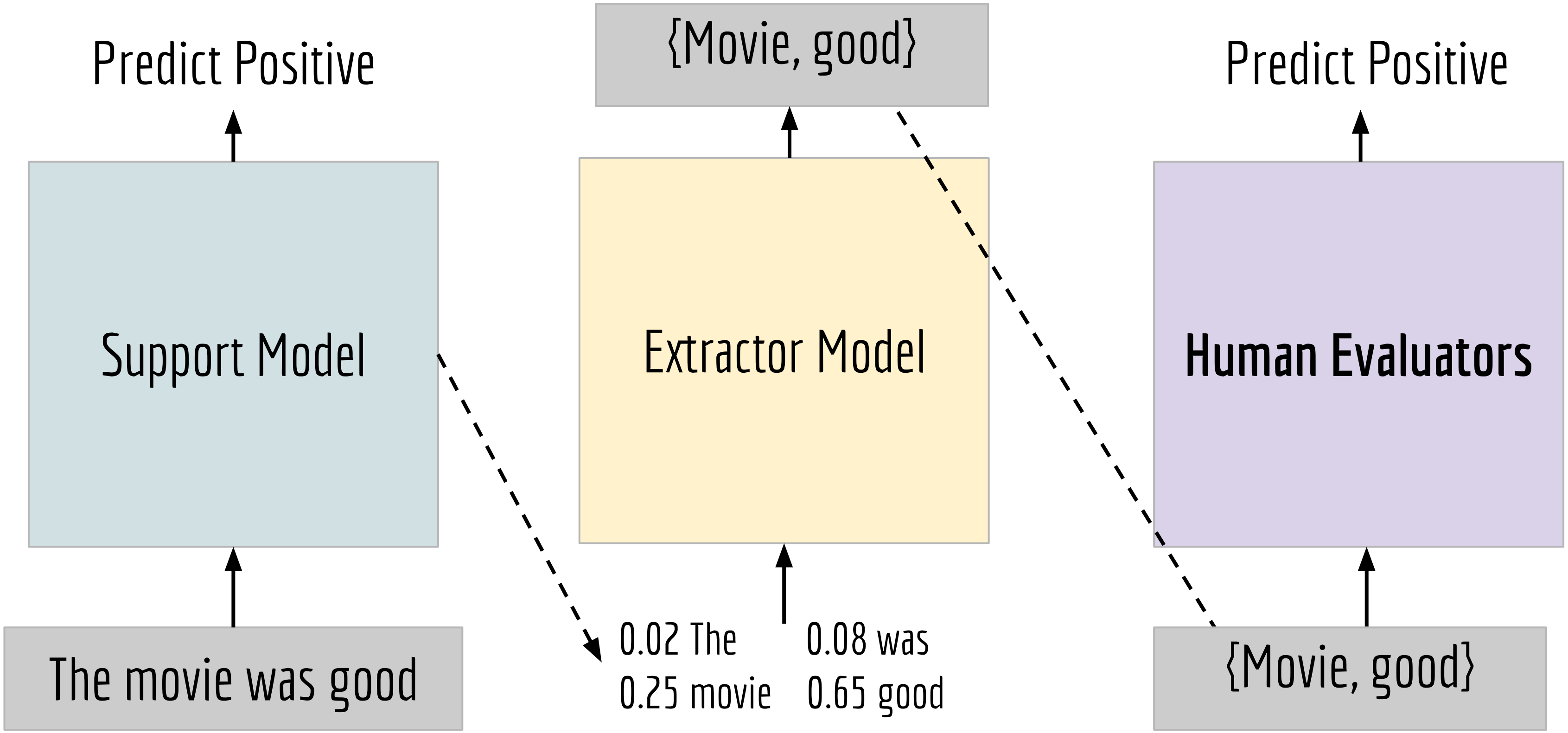
Coherence

- How *readable* and *understandable* are the rationales? (Ehsan et al. 2019, Lei et al. 2016)
- Reflects user preferences

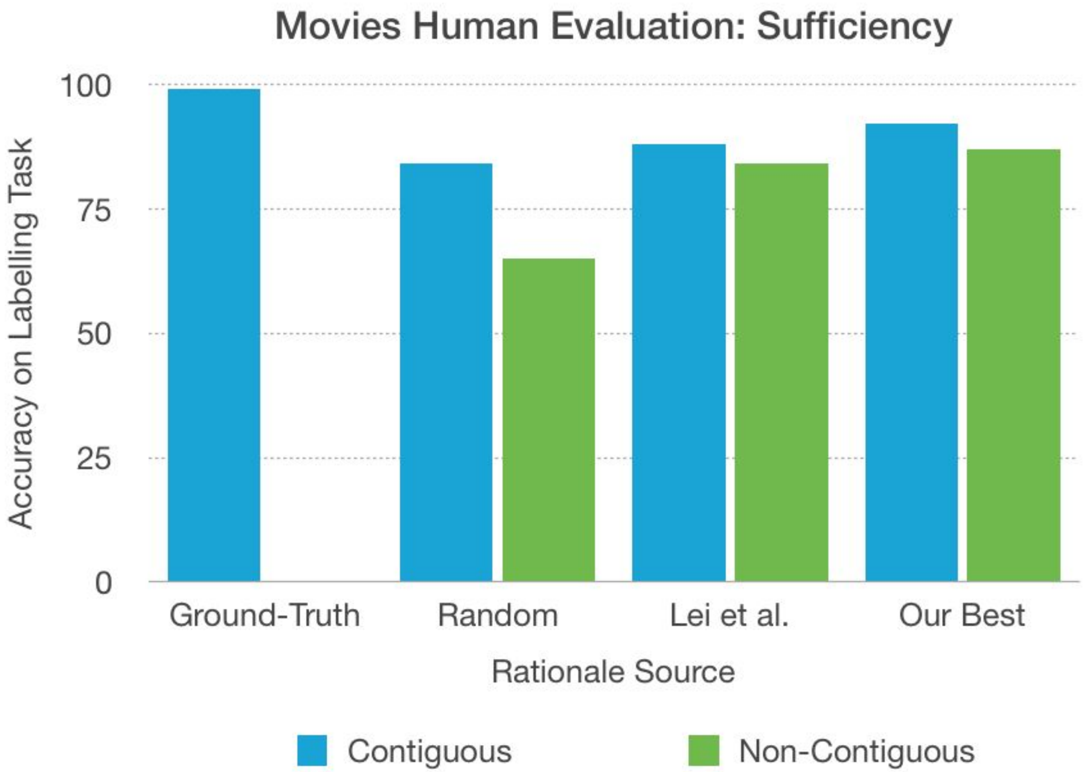
Experiments

- Ask user to perform binary prediction task
    - Movie Reviews: select the sentiment
    - User must perform task *given only the rationale.*
- Ask user to rate their confidence (1-4)
- Ask user to rate the readability (1-5)

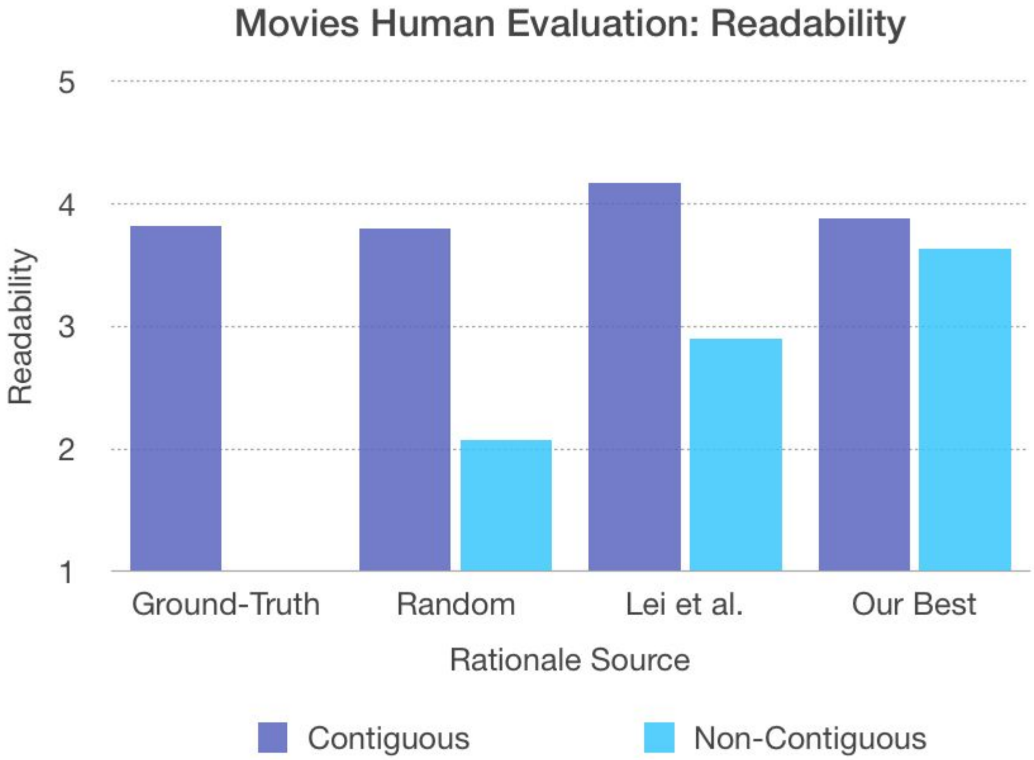
FRESH Model


Predict Positive
Support Model
The movie was good
{Movie, good}
Extractor Model
0.02 The
0.25 movie
0.08 was
0.65 good
Predict Positive
Human Evaluators
{Movie, good}

Results


Movies Human Evaluation: Sufficiency
Accuracy on Labelling Task
100
75
50
25
0
Ground-Truth
Random
Lei et al.
Our Best
Rationale Source
Contiguous
Non-Contiguous

Results



Movies Human Evaluation: Readability
5
4
3
2
1
Readability
Ground-Truth
Random
Lei et al.
Our Best
Rationale Source
Contiguous
Non-Contiguous

Takeaways

1. Terminology matters- it’s important to define what you are looking for.

Takeaways

1. Terminology matters- it’s important to define what you are looking for.
    a. Faithful explainability == model understanding.

Takeaways

1. Terminology matters- it’s important to define what you are looking for.
    a. Faithful explainability == model understanding.
2. No one-size-fits-all answer to “Is Attention (Faithful) Explanation?” debate.

Takeaways

1. Terminology matters- it's important to define what you are looking for.
    a. Faithful explainability == model understanding.
2. No one-size-fits-all answer to "Is Attention (Faithful) Explanation?" debate.
    a. Model components must be tested on a task-specific basis.

Takeaways

1. Terminology matters- it's important to define what you are looking for.
    a. Faithful explainability == model understanding.
2. No one-size-fits-all answer to "Is Attention (Faithful) Explanation?" debate.
    a. Model components must be tested on a task-specific basis.

Takeaways

1. Terminology matters- it’s important to define what you are looking for.
    a. Faithful explainability == model understanding.
2. No one-size-fits-all answer to “Is Attention (Faithful) Explanation?” debate.
    a. Model components must be tested on a task-specific basis.
3. Pipeline approach is one way to guarantee faithfulness (for subset-selection explanations).

Takeaways

1. Terminology matters- it’s important to define what you are looking for.
    a. Faithful explainability == model understanding.
2. No one-size-fits-all answer to “Is Attention (Faithful) Explanation?” debate.
    a. Model components must be tested on a task-specific basis.
3. Pipeline approach is one way to guarantee faithfulness (for subset-selection explanations).
4. Faithfulness and Plausibility are not mutually exclusive criteria.

1. A Foray into Attention Explainability
2. Defining Explanation
3. Is attention explanation?
4. How do we guarantee faithfulness?
5. How do we test plausibility?
6. Future Directions

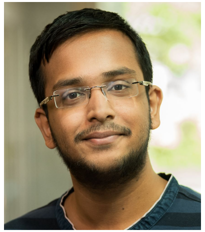
What's Next?

Future Directions

- Model stability & robustness
    - What does variance tell us?
- Better & more consistent human evaluations
- Machine learning approaches to plausibility
    - Leveraging commonsense knowledge/reasoning
- Reinforcement Learning

Thank you!

Collaborators:

Follow me on Twitter:

@sarahwiegreffe

Email: saw@gatech.edu

Code: github.com/sarahwie