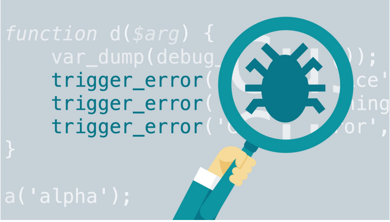
Contrastive Explanations of NLP Models

Ana Marasović

Allen Institute for AI (AI2) × AllenNLP × University of Washington



Technically robust and safe

Allows acknowledging and evaluating trade-offs

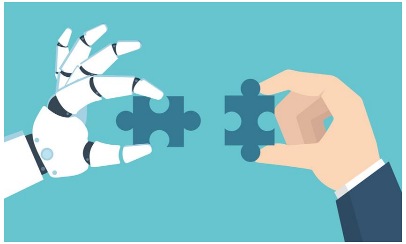
Respects quality and integrity of data

Trustworthy AI Contracts

Encourages green AI

Supports users' agency and oversight

Allows assessing the impact on individuals, society, democracy



European ethics guidelines for trustworthy AI
Jacovi, Marasović, Miller, Goldberg. Formalizing Trust in Artificial Intelligence. FAccT 2021.

One approach to realizing some of the trustworthy AI goals is via **local explanations**: justifications of models' individual predictions

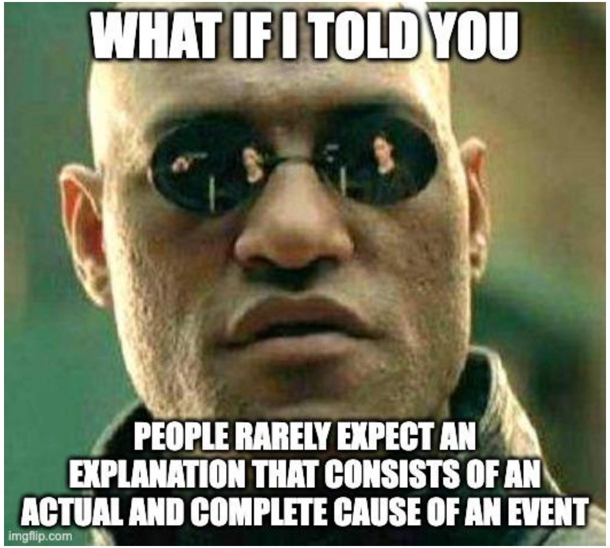
A dominant ML/NLP perspective on local explanations

→ Causal attribution: given a set of factors (usually, input tokens/pixels), select ***all factors*** that ***cause*** the model's decision

Miller. Explanation in artificial intelligence: Insights from the social sciences. In AIJ 2019.

A dominant ML/NLP perspective on local explanations

→ Causal attribution: given a set of factors (usually, input tokens/pixels), select ***all factors*** that ***cause*** the model's decision



Miller's 1st Insight from Social Science

Explanation are **selected (in a biased manner)** because:

1. **Cognitive load**: causal chains are often too large to comprehend

2. Explainee cares only about a small number of causes (relevant to the context)

Miller. Explanation in artificial intelligence: Insights from the social sciences. In AIJ 2019.

Miller's 2nd Insight from Social Science

Explanations are **contrastive** = responses to:

"Why P rather than Q?"

"How to change the answer from P to Q?"

where **P** is an observed event **(fact)**, and **Q** an imagined, counterfactual event that did not occur **(foil)**

Why did she get 20x less limit?

1. Make joint tax returns
2. Live in a community-property state
3. Be married for a long time
4. ....

What are the factors in the application that would need to change to get the same limit?

woman → ?

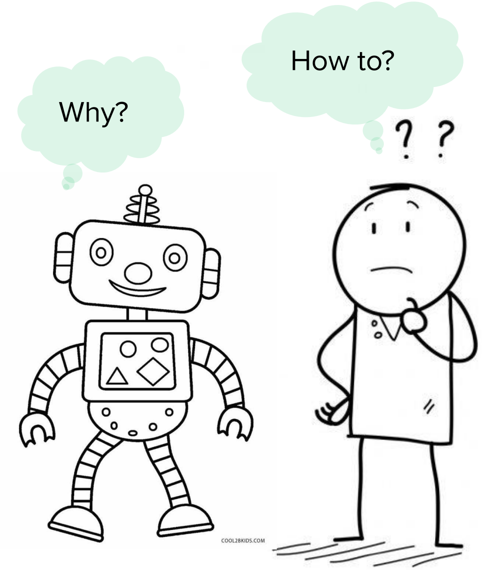
Miller. Explanation in artificial intelligence: Insights from the social sciences. In AIJ 2019.

“Understanding how people define, generate, select, evaluate, and present explanations seems almost essential”

People assign human-like traits to AI models (**anthropomorphic bias**)

⇒ People expect explanations of models’ behavior to follow the same conceptual framework used to explain human behavior

⇒ No users’ agency otherwise



Miller. Explanation in artificial intelligence: Insights from the social sciences. In AIJ 2019.

NLP is starting to pay attention!

COLING 2020 → Yang et al. Generating Plausible **Counterfactual Explanations** for Deep Transformers in Financial Text Classification.

TACL 2021 → Jacovi and Goldberg. Aligning Faithful Interpretations with their Social Attribution.

(Findings of) ACL 2021

→ Chen et al. KACE: Generating Knowledge-Aware **Contrastive Explanations** for NLI.

→ Ross et al. **Explaining** NLP Models via Minimal **Contrastive** Editing (MiCE).

→ Paranjape et al. Prompting **Contrastive Explanations** for Commonsense Reasoning Tasks.

→ Wu et al. Polyjuice: Generating **Counterfactuals** for **Explaining**, Evaluating, and Improving Models

EMNLP 2021 → Jacovi et al. **Contrastive Explanations** for Model Interpretability.

✅ Almost all of these papers begin by citing Miller’s overview of frameworks of explanations from social science

Are technical proposals the same?

Categorization of Current Methods for Contrastive Explanations in NLP

Contrastive Explanations of NLP Models

Contrastive input editing:
Automatic edits to the input that change model output to the contrast case

Yang et al. COLING 2020.

Jacovi and Goldberg. TACL 2020.

Ross et al. Findings of ACL 2021.

Wu et al. ACL 2020.

Collect **free-text** human **contrastive explanations**, ...

...and **generate them** left-to-right Chen et al. ACL 2021.

...abstract them into templates, automatically fill in the templates **(template-based infilling)**

Paranjape et al. Findings of ACL 2021.

Contrastive vector representation:
A dense representation of the input that captures latent features that differentiate two classes

Jacovi et al. EMNLP 2021.

Contrastive Explanations of NLP Models

Contrastive input editing:
Automatic edits to the input that change model output to the contrast case

Yang et al. COLING 2020.

Jacovi and Goldberg. TACL 2020.

Ross et al. Findings of ACL 2021.

Wu et al. ACL 2020.

Collect **free-text** human **contrastive explanations**, …

…and **generate them** left-to-right Chen et al. ACL 2021.

…abstract them into templates, automatically fill in the templates **(template-based infilling)**

Paranjape et al. Findings of ACL 2021.

Contrastive vector representation:
A dense representation of the input that captures latent features that differentiate two classes

Jacovi et al. EMNLP 2021.

Contrastive Explanations via **Contrastive Editing**

The key idea:

"Why P not Q?" ⇒ "How to change the answer from P to Q?"

⇒ By making a **contrastive minimal edit**

A minimal edit to the input that causes the model output to change to the contrast case **has hallmark characteristics of a human contrastive explanation**:

- → cites contrastive features
- → selects a few relevant causes

Ross, Marasović, Peters. MiCE: Explaining NLP Models via Minimal Contrastive Editing. Findings of ACL 2021.

Contrastive Explanations via **Contrastive Editing**

Question:
Ann and her children are going to Linda's home _____.

(a) by bus (b) by car (c) on foot (d) by train

Why **"by train"** (d) and not "**on foot"** (c)?
How to change the answer from **"by train"** (d) to "**on foot"** (c)?

Context:
...Dear Ann, I hope that you and your children will be here in two weeks. My husband and I will go to meet you at the train station. Our town is small...

MiCE-Edited Context:
...Dear Ann, I hope that you and your children will be here in two weeks. My husband and I will go to meet you at ~~the train station~~ **your home on foot**. Our ~~town~~ **house** is small...

Ross, Marasović, Peters. MiCE: Explaining NLP Models via Minimal Contrastive Editing. Findings of ACL 2021.

Contrastive Explanations via **Contrastive Editing**

Question:
Ann and her children are going to Linda's hom _.

(a) by bus (b) by car (c)

Why **"by tra**

How to change

foot" (c)?

I'll go over the details of how to do contrastive editing with MiCE 🐭 later

Context:
...Dear Ann, I hope that you and yo be here in two weeks. My husband go to meet you at the train station. Our own is small...

E-Edited Context:
...Dear Ann, I hope that you and your children will be here in two weeks. My husband and I will go to meet you at ~~the train station~~ **your home on foot**. Our ~~town~~ **house** is small...

Ross, Marasović, Peters. MiCE: Explaining NLP Models via Minimal Contrastive Editing. Findings of ACL 2021.

Contrastive Explanations of NLP Models

Contrastive input editing:
Automatic edits to the input that change model output to the contrast case

Yang et al. COLING 2020.

Jacovi and Goldberg. TACL 2020.

Ross et al. Findings of ACL 2021.

Wu et al. ACL 2020.

Collect **free-text** human **contrastive explanations**, ...

...and **generate them** left-to-right Chen et al. ACL 2021.

...abstract them into templates, automatically fill in the templates **(template-based infilling)**

Paranjape et al. Findings of ACL 2021.

Contrastive vector representation:
A dense representation of the input that captures latent features that differentiate two classes

Jacovi et al. EMNLP 2021.

Contrastive Explanations via **Conditional Generation**

The key idea (IMO):

Contrastive edits could still **not be <u>immediately</u> understandable** (cognitive load could still be notable)

"Why P not Q?" ⇒ Generate **free-text contrastive explanations**

<u>Example</u>: The model predicts "by train" because the context mentions meeting at "the train station". If the context had said that they will meet at "your home on foot" the prediction would be "on foot".

Chen et al. KACE: Generating Knowledge-Aware Contrastive Explanations for Natural Language Inference. ACL 2021.

Contrastive Explanations via **Conditional Generation**

Step 1: Generate **contrastive edits**

(1.a) Highlight important tokens

(1.b) Replace important tokens with WordNet hypernyms and hyponyms

(1.c) Minimize the loss between the predicted and contrast label for examples in (1b)

(1.d) Minimize the distance between the original and edited examples in (1b)

(1.e) Maximize the diversity of edited examples in (1b)

Step 2: Compose a **free-text contrastive explanation** by generating "Why P" and "Why not Q" explanations from **two supervised models**, given the original instance, the contrastively edited instance (Step 1), and external knowledge

end-to-end

Chen et al. KACE: Generating Knowledge-Aware Contrastive Explanations for Natural Language Inference. ACL 2021.

Contrastive Explanations of NLP Models

Contrastive input editing:
Automatic edits to the input that change model output to the contrast case

Yang et al. COLING 2020.

Jacovi and Goldberg. TACL 2020.

Ross et al. Findings of ACL 2021.

Wu et al. ACL 2020.

Collect **free-text** human **contrastive explanations**, …

…and **generate them** left-to-right Chen et al. ACL 2021.

…abstract them into templates, automatically fill in the templates **(template-based infilling)**

Paranjape et al. Findings of ACL 2021.

Contrastive vector representation:
A dense representation of the input that captures latent features that differentiate two classes

Jacovi et al. EMNLP 2021.

Contrastive Explanations via **Template-Based Infilling**

The key idea (IMO):

"Why P not Q?" ⇒ Develop templates (prompts) to retrieve **"contrastive knowledge"**[*] – a comparison of P and Q along a distinguishing attribute – from a pretrained model

[*] Example: *Peanuts are salty while raisins tend to be sweet.*

Paranjape et al. Prompting Contrastive Explanations for Commonsense Reasoning Tasks. Findings of ACL 2021.

Contrastive Explanations via **Template-Based Infilling**

How to tie pieces of paper together?
(a) Thread **ruler** through the holes.
(b) Thread **ribbon** through the holes. [correct]

Data Step 1: Collect **human-written** free-text contrastive explanations

Human contrastive explanation:
Ruler ***is*** hard ***while*** a **ribbon** ***is*** flexible.

Data Step 2: Abstract them into templates with placeholders

Template:
P ***is*** ____ ***while*** **Q** ***is*** ____

Paranjape et al. Prompting Contrastive Explanations for Commonsense Reasoning Tasks. Findings of ACL 2021.

Contrastive Explanations via **Template-Based Infilling**

Modeling Step 1:

Generate contrastive knowledge by filling in the placeholders in explanation templates

Templates:
→ P **is ___ while** Q **is ___**
→ P **takes longer to ___ that** Q
→ P **can cause ___ while** Q **results in ___**
...

To prepare the puff pastry for you pie, line a baking sheet with parchment. Then ___
(a) Unroll the pastry, lay it over **baking twine**. [correct]
(b) Unroll the pastry, lay it over **fishing line**.

Contrastive knowledge:
→ **Baking twine is** used in baking **while fishing line is** used in fishing.
→ **Baking twine takes longer to** catch fish **than fishing line.**
→ **Baking twine can cause** fire **while fishing line results in** tangling.
...

Paranjape et al. Prompting Contrastive Explanations for Commonsense Reasoning Tasks. Findings of ACL 2021.

Contrastive Explanations via **Template-Based Infilling**

Modeling Step 2:

Augment the input with contrastive knowledge and make a prediction with the same model

To prepare the puff pastry for you pie, line a baking sheet with parchment. Then unroll the pastry, lay it over baking twine.

To prepare the puff pastry for you pie, line a baking sheet with parchment. Then unroll the pastry, lay it over **fishing line**.

×

Contrastive knowledge:

→ **Baking twine is** used in baking **while fishing line is** used in fishing.

→ **Baking twine takes longer to** catch fish **than fishing line.**

→ **Baking twine can cause** fire **while fishing line** results in tangling.

...

model scores (context c, answer candidate a_i, contrastive knowledge e_j) tuples

⬇

$$\text{argmax}_i \sum_j \text{score}(c, a_i, e_j)$$

⬇

The highest scoring explanation is THE explanation

Paranjape et al. Prompting Contrastive Explanations for Commonsense Reasoning Tasks. Findings of ACL 2021.

Contrastive Explanations of NLP Models

Contrastive input editing:
Automatic edits to the input that change model output to the contrast case

Yang et al. COLING 2020.

Jacovi and Goldberg. TACL 2020.

Ross et al. Findings of ACL 2021.

Wu et al. ACL 2020.

Collect **free-text** human **contrastive explanations**, …

…and **generate them** left-to-right Chen et al. ACL 2021.

…abstract them into templates, automatically fill in the templates **(template-based infilling)**

Paranjape et al. Findings of ACL 2021.

Contrastive vector representation:
A dense representation of the input that captures latent features that differentiate two classes

Jacovi et al. EMNLP 2021.

Contrastive Explanations via **Contrastive Projection**

The key idea (IMO):

"Why P not Q?" ⇒ Select **latent** contrastive features in the space of **hidden representations** instead of selecting them in the input (discrete tokens)

Jacovi et al. Contrastive Explanations for Model Interpretability. EMNLP 2021.

Contrastive Explanations via **Contrastive Projection**

Thesis: Entailment because of a high lexical overlap between the premise and hypothesis

***Overlap concept*:** All of the content words in the hypothesis also exist in the premise

Causal Intervention (Why P?)

→ Study how model logits change by removing all features in the hidden representation indicative of the overlap concept

Jacovi et al. Contrastive Explanations for Model Interpretability. EMNLP 2021.

Contrastive Explanations via **Contrastive Projection**

Thesis: Entailment because of a high lexical overlap between the premise and hypothesis

***Overlap concept*:** All of the content words in the hypothesis also exist in the premise

Causal Intervention (Why P?)

→ Study how model logits change by removing all features in the hidden representation indicative of the overlap concept

Doesn't answer if (a subset of) these features differentiate entailment from other classes

Jacovi et al. Contrastive Explanations for Model Interpretability. EMNLP 2021.

Contrastive Explanations via **Contrastive Projection**

Thesis: Entailment because of a high lexical overlap between the premise and hypothesis

Overlap concept**:** All of the content words in the hypothesis also exist in the premise

Contrastive Intervention (Why P not Q?)

→ Project the hidden representation to the space of contrastive feature, i.e., **remove hidden features** that the **model doesn't use to differentiate class P (entailment) from class Q (contradiction or neutral)**

→ Study how model logits change by **removing all features in the contrastively projected hidden representation** indicative of the overlap concept

Jacovi et al. Contrastive Explanations for Model Interpretability. EMNLP 2021.

✅ NLP is starting to acknowledge the perspective of the social sciences on explainability

💡 Proposed methods for producing contrastive explanations differ:

1. Contrastive editing
2. Free-text contrastive explanations
3. Contrastive vector representations

❗ Specify what kind of contrastive explanations you aim to build

Deeper Into Contrastive Editing

Alexis Ross, Ana Marasović, Matt Peters (2021)

Explaining NLP Models via Minimal Contrastive Editing (MiCE)

Goal:

Automatically find a **minimal edit** to the input that **causes the model output to change to the contrast case**

A very high-level idea of 🐭:

Keep masking and filling masked positions until you find an edit that flips the label, while simultaneously minimizing the masking percentage (i.e., the edit size)

the contrast label (foil)

```
label: positive input: Sylvester Stallone has made some crap films in his lifetime,
but this has got to be one of the worst. A totally dull story...
```

the contrast label (foil)

`label: positive input: Sylvester Stallone has made some crap films in his lifetime, but this has got to be one of the worst. A totally dull story...`

mask *n*% of input tokens

label: positive input: Sylvester Stallone has made some **<mask>** films in his lifetime, but this has got to be one of the **<mask>**. A totally **<mask>** story...

the contrast label (foil)

`label: positive input: Sylvester Stallone has made some crap films in his lifetime, but this has got to be one of the worst. A totally dull story...`

mask *n*% of input tokens

label: positive input: Sylvester Stallone has made some **<mask>** films in his lifetime, but this has got to be one of the **<mask>**. A totally **<mask>** story...

sample *m* spans at each masked position

1. label: positive input: Sylvester Stallone has made some **good** films in his lifetime, but this has got to be one of the **worst**. A totally **novel** story...

2. label: positive input: Sylvester Stallone has made some **great** films in his lifetime, but this has got to be one of the **greatest of all time**. A totally **boring** story...

...

m. label: positive input: Sylvester Stallone has made some **wonderful** films in his lifetime, but this has got to be one of the **greatest**. A totally **tedious** story...

the contrast label (foil)

`label: positive input: Sylvester Stallone has made some crap films in his lifetime, but this has got to be one of the worst. A totally dull story...`

mask *n*% of input tokens

label: positive input: Sylvester Stallone has made some **<mask>** films in his lifetime, but this has got to be one of the **<mask>**. A totally **<mask>** story...

sample *m* spans at each masked position

get the probability of the contrast label

1. label: positive input: Sylvester Stallone has made some **good** films in his lifetime, but this has got to be one of the **worst**. A totally **novel** story... $\mathbb{P}(pos) = 0.2$

2. label: positive input: Sylvester Stallone has made some **great** films in his lifetime, but this has got to be one of the **greatest of all time**. A totally **boring** story... $\mathbb{P}(pos) = 0.6$

...

m. label: positive input: Sylvester Stallone has made some **wonderful** films in his lifetime, but this has got to be one of the **greatest**. A totally **tedious** story... $\mathbb{P}(pos) = 0.65$

1. Prepend the contrast label to the input
2. Mask *n*% of the input tokens
3. Sample *m* spans at masked positions

×

s **different values of *n* to minimize the edit***

* s=4 in the paper

How to pick which values for *n*?

Binary search on [0,55]

1. Prepend the contrast label to the input
2. Mask n% of the input tokens
3. Sample m spans at masked positions

$\times$

s different values of n to minimize the edit*

* s=4 in the paper

How to pick which values for n?

Binary search on [0,55]

Start: $n^{(1)}$=27.5%

1. Prepend the contrast label to the input
2. Mask *n*% of the input tokens
3. Sample *m* spans at masked positions

$\times$

s* different values of *n* to minimize the edit

* s=4 in the paper

How to pick which values for *n*?

Binary search on [0,55]

Start: $n^{(1)}$=27.5%

➔ If a contrastive edit found: $n^{(2)}$=13.75%

1. Prepend the contrast label to the input
2. Mask n% of the input tokens
3. Sample m spans at masked positions

$\times$

s* different values of *n* to minimize the edit

* s=4 in the paper

How to pick which values for *n*?

Binary search on [0,55]

Start: $n^{(1)}=27.5\%$

➔ If a contrastive edit found: $n^{(2)}=13.75\%$

➔ If a contrastive edit **not** found: $n^{(2)}=41.25\%$

1. Prepend the contrast label to the input
2. Mask *n*% of the input tokens
3. Sample *m* spans at masked positions

× ***s* different values of *n* to minimize the edit***

* s=4 in the paper

How to pick which values for *n*?

Binary search on [0,55]

Start: $n^{(1)}$=27.5%

→ If a contrastive edit found: $n^{(2)}$=13.75%
- If a contrastive edit found: $n^{(3)}$=6.875%

→ If a contrastive edit **not** found: $n^{(2)}$=41.25%
- If a contrastive edit found: $n^{(3)}$=20.625%

1. Prepend the contrast label to the input
2. Mask n% of the input tokens
3. Sample m spans at masked positions

$\times$

s* different values of *n* to minimize the edit

* s=4 in the paper

How to pick which values for *n*?

Binary search on [0,55]

Start: $n^{(1)}$=27.5%

- If a contrastive edit found: $n^{(2)}$=13.75%
  - If a contrastive edit found: $n^{(3)}$=6.875%
  - If a contrastive edit **not** found: $n^{(3)}$=20.625%
- If a contrastive edit **not** found: $n^{(2)}$=41.25%
  - If a contrastive edit found: $n^{(3)}$=20.625%
  - If a contrastive edit **not** found: $n^{(3)}$=48.125%

1. Prepend the contrast label to the input
2. Mask *n*% of the input tokens
3. Sample *m* spans at masked positions

× ***s* different values of *n* to minimize the edit***

* s=4 in the paper

How to pick masking positions?

Based on token importance for the original prediction

Rank input tokens based on the magnitude of the gradients of the model we're explaining

Mask top-*n*% of **ranked** tokens

We find that this works better than randomly masking tokens

1. Prepend the contrast label to the input
2. Mask *n%* of the input tokens
3. Sample m spans at masked positions

× ***s* different values of *n* to minimize the edit***

* s=4 in the paper

$s*m$ samples

* m=15 in the paper

1. Prepend the contrast label to the input
2. Mask n% of the input tokens
3. Sample m spans at masked positions

$\times$

s different values of n to minimize the edit*

* s=4 in the paper

$s*m$ samples

* m=15 in the paper

rank $s*m$ samples w.r.t. the probability of the contrast label

1. Prepend the contrast label to the input
2. Mask *n%* of the input tokens
3. Sample *m* spans at masked positions

×

s* different values of *n* to minimize the edit

* s=4 in the paper

$s*m$ samples

* m=15 in the paper

rank $s*m$ samples w.r.t. the probability of the contrast label

beam keep top-*b* samples

* b=3 in the paper

1. Prepend the contrast label to the input
2. Mask *n%* of the input tokens
3. Sample *m* spans at masked positions

× ***s* different values of *n* to minimize the edit***

* s=4 in the paper

s∗*m* samples

* m=15 in the paper

rank *s*∗*m* samples w.r.t. the probability of the contrast label

beam keep top-*b* samples

* b=3 in the paper

if the contrastive edit is found

repeat these steps for every instance in the beam for 2 more rounds

1. Prepend the contrast label to the input
2. Mask *n*% of the input tokens
3. Sample *m* spans at masked positions

× ***s* different values of *n* to minimize the edit***

* s=4 in the paper

*s*m* samples

* m=15 in the paper

rank *s*m* samples w.r.t. the probability of the contrast label

beam keep top-*b* samples

* b=3 in the paper

The maximum number of iterations for a single instance:

$$\overbrace{\text{\# binary search levels } \mathtt{s} \times \text{\# samples at each maskin position } \mathtt{m}}^{\text{first round}} +$$

$$\underbrace{\text{beam size } \mathtt{b} \times \text{\# binary search levels } \mathtt{s} \times \text{\# samples at each masking position } \mathtt{m} \times \text{\# of rounds}}_{\text{other rounds}} =$$

$4 \times 15 + 3 \times 4 \times 15 \times 2 = 420$

Can a pretrained model without any additional tweaks fill in the spans?

So-so

We find that **preparing the editor** by finetuning it to infill masked spans given masked text and **a target end-task label** as input is an important step before using it to editing

Can a pretrained model without any additional tweaks fill in the spans?

So-so

We find that **preparing the editor** by finetuning it to infill masked spans given masked text and **a target end-task label** as input is an important step before using it to editing

We find that **labels predicted by the model** we're explaining **can be used** in this step without a big loss in performance (good option if you don't have the labeled data)

Can a pretrained model without any additional tweaks fill in the spans?

So-so

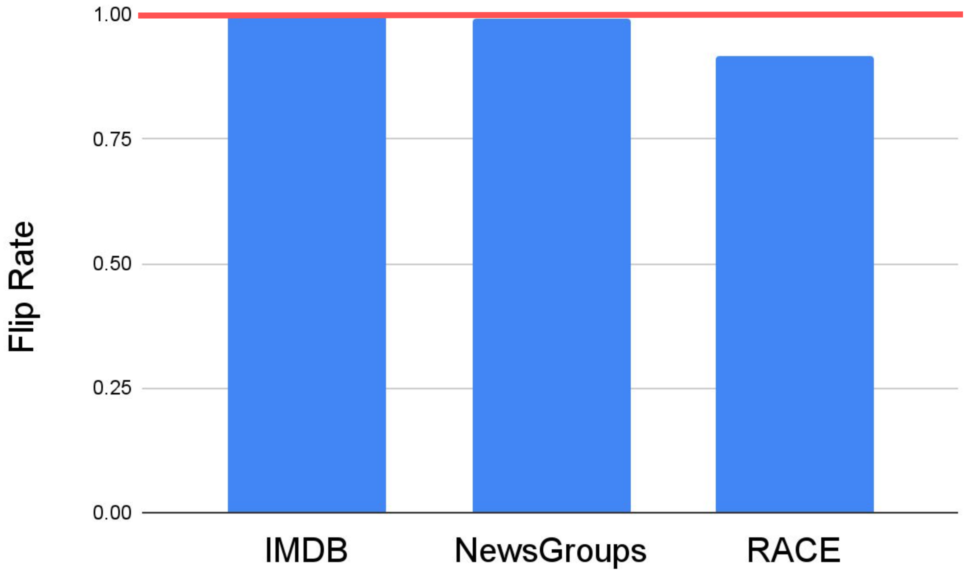
We find that **preparing the editor** by finetuning it to infill masked spans given masked text and **a target end-task label** as input is an important step before using it to editing

We find that **labels predicted by the model** we're explaining **can be used** in this step without a big loss in performance (good option if you don't have the labeled data)

⇒ **MiCE is a two-stage approach** to generating contrastive edits

Stage 1: prepare an editor
Stage 2: make edits using the editor guided by the gradients and logits of the predictor we're explaining

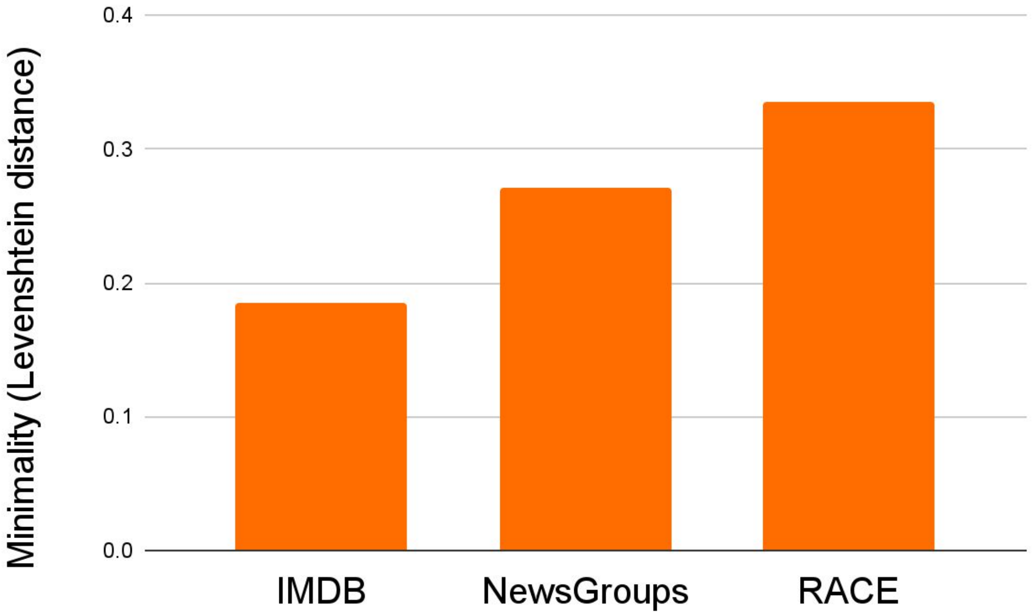
Results – Flip Rate



1.0 when we find a contrastive edit for all instances

Results – Edit Minimality

The minimum number of deletions, insertions, or substitutions required to transform the original to the
edited instance

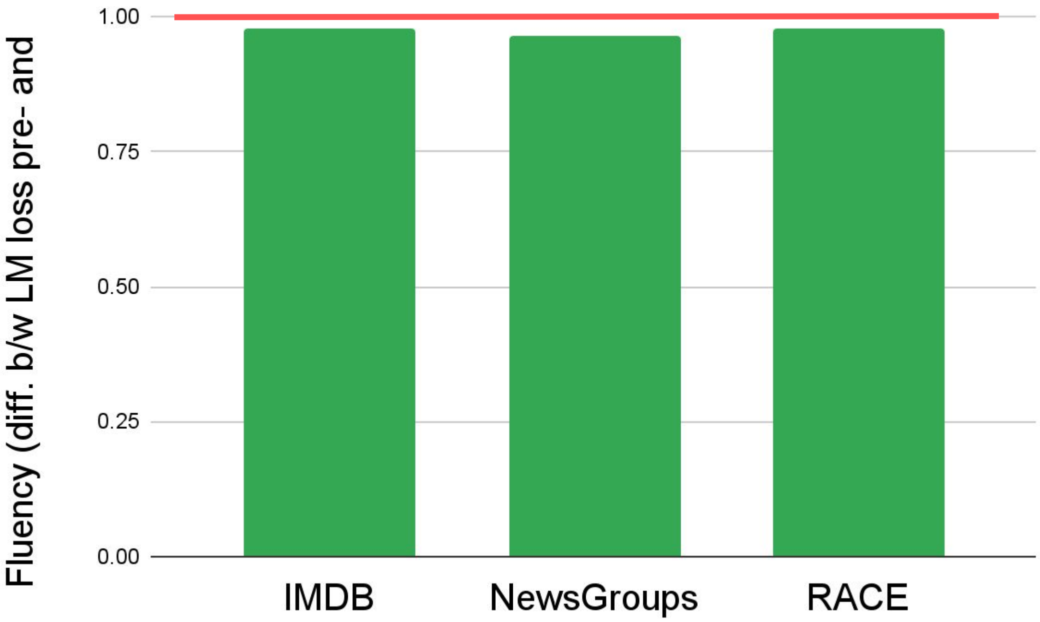


lower is better; we change on average 18.5-33.5% of the input tokens

The size of the IMDB edits is similar to human edits*

* Compared to IMDB edits in Gardner et al. Evaluating Models' Local Decision Boundaries via Contrast Sets. Findings of EMNLP 2020.

Results – Edit Fluency



1.0 when a LM loss pre- and post-editing doesn't change

How Can MiCE Edits Be Used?

MiCE's edits can offer hypotheses about model "bugs"

Original pred $y_p = \underline{\text{positive}}$ **Contrast pred** $y_c = \text{negative}$

An interesting pairing of stories, this little flick manages to bring together seemingly different characters and story lines all in the backdrop of WWII and succeeds in tying them together without losing the audience. I was impressed by the depth portrayed by the different characters and also by how much I really felt I understood them and their motivations, even though the time spent on the development of each character was very limited. The outstanding acting abilities of the individuals involved with this picture are easily noted. A fun, stylized movie with a slew of comic moments and a bunch more head shaking events. ~~7/10~~ **4/10**

How Can MiCE Edits Be Used?

MiCE's edits can offer hypotheses about model "bugs"

Hypothesis: Model learned to rely heavily on numerical ratings ⭐

Test the hypothesis using MiCE's edits:

1. Filter instances for which the MiCE edit has a minimality value of ≤ 0.05

2. Select tokens that are removed/inserted at a higher rate than expected given the frequency with which they appear in the original IMDB inputs

$y_c =$ ***positive***		$y_c =$ ***negative***	
Removed	**Inserted**	**Removed**	**Inserted**
4/10	excellent	10/10	awful
ridiculous	enjoy	8/10	disappointed
horrible	amazing	7/10	1
4	entertaining	9	4
predictable	10	enjoyable	annoying

Want to know more about MiCE?

Alexis is presenting a poster at BlackboxNLP!

✅ NLP is starting to acknowledge the perspective of the social sciences on explainability

💡 Proposed methods for producing contrastive explanations differ:

1. Contrastive editing
2. Free-text contrastive explanations
3. Contrastive vector representations

❗ Specify what kind of contrastive explanations you aim to build

✅ **Contrastive editing is already achieving decent performance**

❗ **Obviously needed improvements: less iterations & more precise minimality**

(Contrastive) Local Explanations: What is Next?

Miller's 1st Insight from Social Science

Explanation are **selected (in a biased manner)** because:

1. **Cognitive load**: causal chains are often too large to comprehend

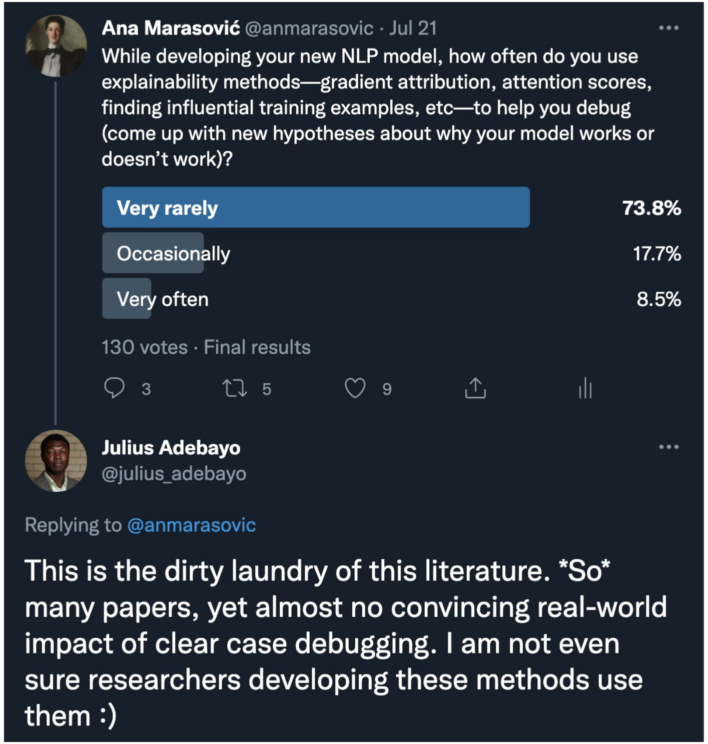
2. Explainee cares only about a small number of causes (relevant to the context)

We don't test whether generated contrastive explanations are more easily understood or whether they match people's expectations

Miller. Explanation in artificial intelligence: Insights from the social sciences. In AIJ 2019.

This is not specific to contrastive explanations…

Although local explanations are specifically motivated for people to use, there is no convincing evidence that local explanations help people who are using language technology



We Lack Evidence That Local Explanations Are Helpful

This is in part due to:

- **Focus on grand AI challenges**, but not useful applications
- **Simple tasks** that people don't need help with (e.g., commonsense QA)
- The use of automatic measures of explanation plausibility **without specifying what real-world situations highly plausible explanations will help with**

We Lack Evidence That Local Explanations Are Helpful

This is in part due to:

- **Focus on grand AI challenges**, but not useful applications
- **Simple tasks** that people don't need help with (e.g., commonsense QA)
- The use of automatic measures of explanation plausibility **without specifying what real-world situations highly plausible explanations will help with**

To meaningfully move forward we need to answer:

➔ **What are potentially useful language applications and who is targeted audience?**
(e.g., journalist and fact checking)

➔ **How explanations might help people using these applications?**
(e.g., by helping them verify information faster without the loss of accuracy)

➔ **Test them exactly for those purposes**

Thank you!

Miller's 4th Insights from Social Science

Explanations are social: we interact and argue about the explanation and contextualize explanation w.r.t. the explainee

Why is image J labelled as a Spider instead of a Beetle?

Because the arthropod in image J has 8 legs, consistent with those in the category Spider, while those in Beetle has 6 legs.

Why did you infer that the arthropod in image J has 8 legs instead of 6?

I counted the 8 legs that I found, as I have just highlighted on the image now.

Miller. Explanation in artificial intelligence: Insights from the social sciences. In AIJ 2019.