Situated Natural Language Explanations

Zining Zhu, Haoming Jiang, Jingfeng Yang, Sreyashi Nag, Chao Zhang,
Jie Huang, Yifan Gao, Frank Rudzicz, Bing Yin

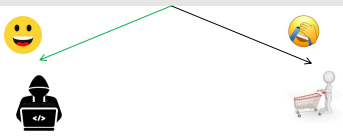
Explanation should be situated

Natural language explanations have the potentials to communicate complex decisions to a wide variety of audience, but the quality of explanations are not static.

Here's an examples of two explanations with different perceived qualities for different audience:

I searched for "bike" and the app recommended a helmet --

Because the app searched in a database for the product with the highest vector similarity to the query, and found this helmet.



Audience 1: AI engineer
- Cares about the details.
- Cares about mechanisms.

Audience 2: Casual customer
- Randomly shops around.
- Doesn't spend more than 1 second.

Because wearing a helmet makes bike riding safe.

A convenient framework to generate situated NLEs

We propose a framework to adapt the generation of natural language explanations towards the situations of the audience.

Step 1
Identify a desired property of explanation in the situation.

Explain to
Shorter is better.

Step 2
Prompt a PLM using a collection of *situated* prompts.

[…] because, in short
[…] because, basically
[…] because, essentially
I'm busy. […] because
…

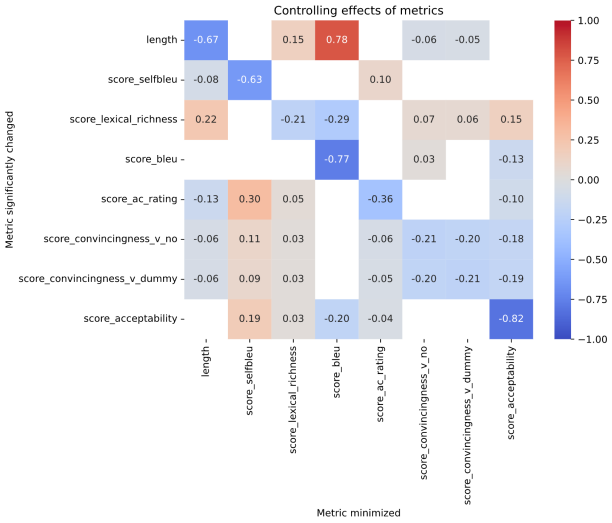
Step 3
Select the generated explanations using the identified property.

[…] because, in short, wearing a helmet makes bike riding safe.

Situated prompt → Situated explanation

Automatic evaluation of explanation quality

Automatic evaluations describe the explanation quality along multiple dimensions, and reveal the strategies that PLMs adopt to generate the explanations.



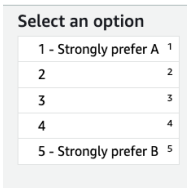
Human evaluation of situated explanation

We simulate situations by giving different time constraints and ask annotators to select their preferences towards different situated explanations.

I searched for "bike" and the app recommended a helmet --

Explanation A is generated with the hint "in short" in the prompt.
Explanation B is generated with the hint "in detail" in the prompt.

Which explanation better explains the recommendation?



Take your time and read the texts carefully.

Preference = 3.10 (sd = 0.73)

Try to finish this question in 20 seconds.

Preference = 2.98 (sd = 0.74)

Given simulated scenarios, human annotators significantly ($p = 0.038$) prefer the explanations correctly adapted to their situations. There are other evaluations in the paper.

Takeaways

Automatic generation of natural language explanations should be situated towards the audience, and we propose a convenient framework to do so.