Case-Based Reasoning with Language Models for Classification of Logical Fallacies

USC Viterbi School of Engineering Information Sciences Institute

Zhivar Sourati[1,2], Filip Ilievski[1,2], Hong- An Sandlin[3], and Alain Mermoud[3]
[1]Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA
[2]Department of Computer Science, University of Southern California, Los Angeles, CA, USA
[3]Cyber-Defence Campus, armasuisse Science and Technology, Switzerland
{souratih,ilievski}@isi.edu, {hongan.sandlin,alain.mermoud}@ar.admin.ch

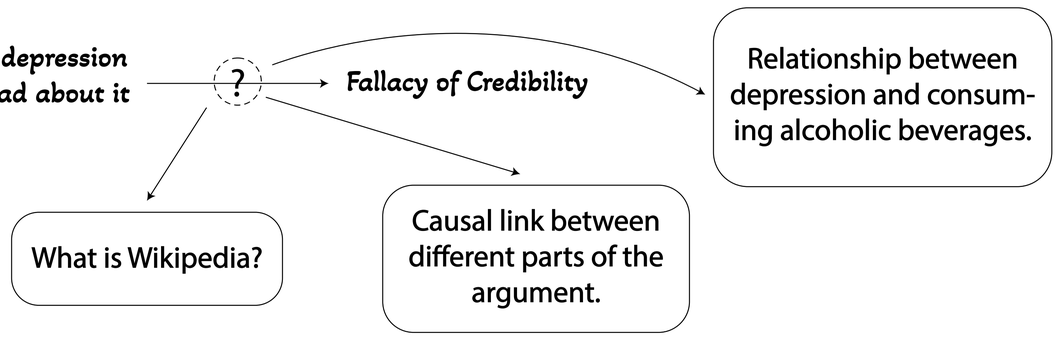
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
armasuisse Science and Technology

Motivation

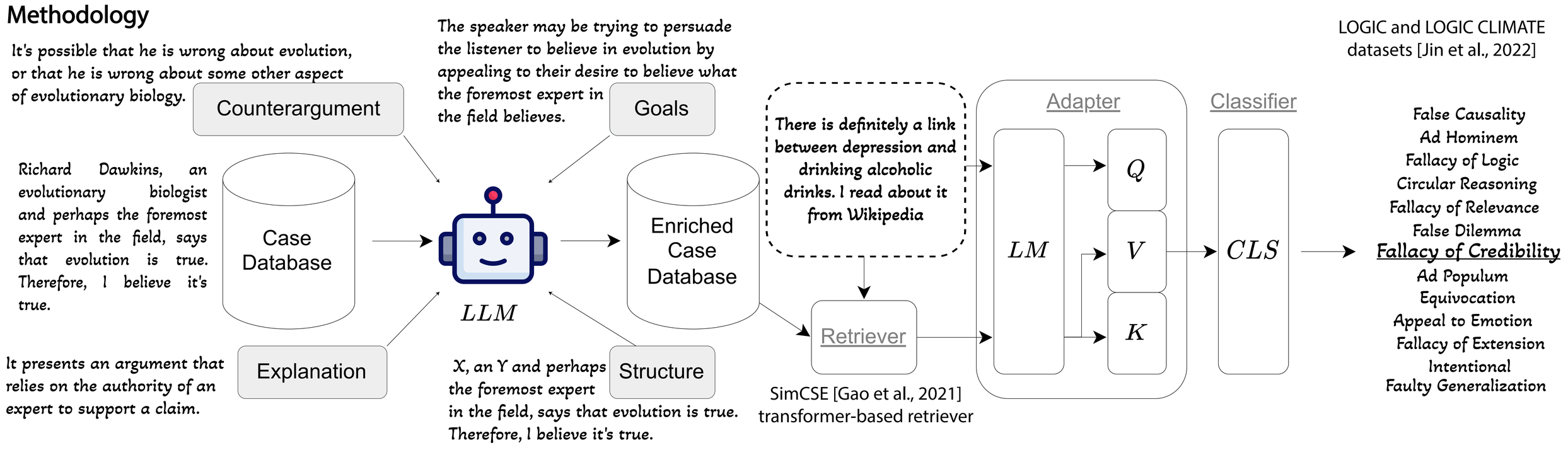
Logical fallacies are arguments that may sound convincing but are based on faulty logic and are therefore invalid. Classifying logical fallacies is hard both for humans and language models (LMs).

Case-based reasoning (CBR) solves new problems based on the solutions of similar past problems by **retrieving** prior similar cases and **adapting** them to the current situation.

RQ1: Does reasoning over examples improve the ability of language models to classify logical fallacies?
RQ2: How does case representation affect Case-based reasoning on logical fallacies?



Methodology



Results

Model	Type	LOGIC P	LOGIC R	LOGIC F1	LOGIC Climate P	LOGIC Climate R	LOGIC Climate F1
Freq-based	baseline	0.094	0.094	0.093	0.120	0.079	0.080
Codex	few-shot	0.594	0.422	0.386	0.198	0.093	0.077
ELECTRA	baseline	0.614	0.602	0.599	0.276	0.229	0.217
	CBR	**0.663**	**0.664**	**0.657**	**0.355**	**0.254**	**0.270**
RoBERTa	baseline	0.577	0.561	0.560	0.237	0.211	0.200
	CBR	**0.631**	**0.619**	**0.619**	**0.379**	**0.248**	**0.245**
BERT	baseline	0.585	0.598	0.586	0.166	0.130	0.120
	CBR	**0.613**	**0.616**	**0.611**	**0.359**	**0.204**	**0.200**

Overall, the CBR method brings a consistent improvement in the classification of logical fallacies by LMs.

CBR improves the performance on the out-of-domain benchmark (LOGIC Climate) as well as on the classes with limited training data.

Counterarguments that anticipate and remove any doubts about arguments and structural views that demonstrate an abstraction over the argument help the most among case representations.

Model	Representation	LOGIC P	LOGIC R	LOGIC F1	LOGIC Climate P	LOGIC Climate R	LOGIC Climate F1
ELECTRA	*Text*	0.655	0.634	0.635	0.317	0.242	0.242
	Counterarg.	**0.663**	**0.664**	**0.657**	0.355	**0.254**	**0.270**
	Goals	0.646	0.622	0.621	**0.376**	0.217	0.222
	Structure	0.634	0.625	0.618	0.375	0.254	0.269
	Explanations	0.605	0.580	0.578	0.314	0.242	0.237
RoBERTa	Text	**0.633**	0.613	0.619	0.343	0.236	0.251
	Counterarg.	0.624	0.613	0.615	0.367	0.198	0.216
	Goals	0.632	0.613	0.619	0.351	0.242	**0.263**
	Structure	0.631	**0.619**	**0.619**	**0.379**	**0.248**	0.245
	Explanations	0.575	0.558	0.559	0.359	0.192	0.181
BERT	*Text*	0.595	0.604	0.596	0.311	0.192	0.204
	Counterarg.	0.607	0.613	0.603	0.342	**0.217**	**0.228**
	Goals	0.598	0.607	0.596	0.310	0.204	0.203
	Structure	**0.613**	**0.616**	**0.611**	**0.359**	0.204	0.200
	Explanations	0.540	0.531	0.532	0.274	0.217	0.190

Case Study

	Input Sentence	Enriched Representation for Correct Prediction (representation)	Enriched Representation for Wrong Prediction (representation) (predicted class)	Class
1.	People who don't support the proposed minimum wage increase hate the poor.	There are often multiple perspectives on an issue. It's possible to have a nuanced or balanced view that doesn't align with any side completely. ***(Counterarg.)***	That candidate wants to raise the minimum wage, but they aren't even smart enough to run a business. ***(Text) (Ad Hominem)***	*Fallacy of Extension*
2.	The house is white; therefore it must be big.	X is y; therefore, it is z. ***(Structure)***	The sentence "People who drive big cars hate the environment" presents a generalization about a group of people without sufficient evidence and it relies on oversimplification.***(Explanations) (Faulty Generalization)***	*Fallacy of Logic*
3.	Student: You didn't teach us this; we never learned this. Teacher: So, you're either lazy or unwilling to learn is that right?	It's possible that the argument "It's possible to pass the class without attending. so, you will pass even if you don't attend" is trying to convince the listener that they will pass the class even if they don't attend. The speaker may be trying to persuade the listener to skip class. ***(Goals)***	The sentence "Teacher: You are receiving a zero because you didn't do your homework. Students: Are you serious? You gave me a zero because you hate me?" attacks the person making the argument rather than the argument itself. ***(Explanations) (Fallacy of Extension)***	*False Dilemma*
4.	One day, Megan wore a Donald Duck shirt, and she got an A on her test. Now she wears that shirt every day to class.	There are many factors that contribute to a student's grade, and it's not fair to suggest that the student's past grades are the only factor. It's possible that the student failed the test because they didn't study, or because they were sick. ***(Counterarg.)***	The sentence "Eating five candy bars and drinking two sodas before a test helps me get better grades. I did that and got an A on my last test in history" presents a causal relationship between two events without sufficient evidence to support the claim. ***(Explanations) (Fallacy of Relevance)***	*False Causality*

Representation	LOGIC ground truth overlap	LOGIC predictions overlap	LOGIC Climate ground truth overlap	LOGIC Climate predictions overlap
Text	0.184	0.232	**0.136**	0.173
Counterarg.	0.208	0.220	0.062	0.068
Goals	0.178	0.196	0.130	0.124
Structure	0.238	0.250	0.105	0.242
Explanations	**0.277**	**0.447**	0.086	**0.478**

Overlap of retrieved cases' labels with true labels and predictions of the best CBR model (ELECTRA). We highlight the highest overlaps in **bold**.

Applying the same label as the retrieved cases on the new cases, like K-nearest neighbors (KNN), shows poor performance.

The model that has the highest direct effect on the predictions (*Explanations* representation) shows suboptimal performance.

1. Retrieved case helps the model indirectly by providing CBR with high-level counterarguments, while similar case with high surface level similarity can be confusing for the model.
2. Symbolic abstraction that is achieved by extracting and retrieving a similar case by its structure is helpful for the model.
3. Writer's goal expressed in an explicit manner can be more helpful than a simple general explanation over the argument.
4. Alternative possibilities, formalized as counterarguments is the most helpful enrichment as can be in seen in this example and in general, both in in-domain and out-of-domain settings.

SCAN ME

github.com/zhpinkman/CBR