KoCoSa: Korean Context-aware Sarcasm Detection Dataset

Yumin Kim*, Heejae Suh*, Mingi Kim*, Dongyeon Won*, Hwanhee Lee

Language Intelligence lab, Chung-Ang University

Context-aware Sarcasm Detection

- Sarcasm is a form of verbal irony characterized by saying something contrary to the text's literal meaning. Therefore providing context is essential in sarcasm detection.

Input: A: 아, 당신의 영화 선택은 정말 훌륭한가 보네요 } *Target Utterance*
(A: Oh, your movie choices are just so brilliant, aren't they?)

↓ Sarcasm Detection Model

No Sarcasm

(A) Sarcasm detection without context

Input:
B: 영화도 보고, 뮤지컬도 보러 가고 그래요. 많이 즐기고 있습니다.
A: 그렇군요. 무슨 영화를 보셨나요?
B: 그게, 사실 제가 본 영화들이 대부분 너무 재미없더라고요. } *Context*
A: 아, 당신의 영화 선택은 정말 훌륭한가 보네요 } *Target Utterance*
(B: Go watch movies and go to the musical. I'm enjoying it a lot.
A: I see. What movie did you watch?
B: Actually, most of the movies I watched were so boring.
A: Oh, your movie choices are just so brilliant, aren't they?)

↓ Sarcasm Detection Model

Sarcasm

(B) Sarcasm detection with context

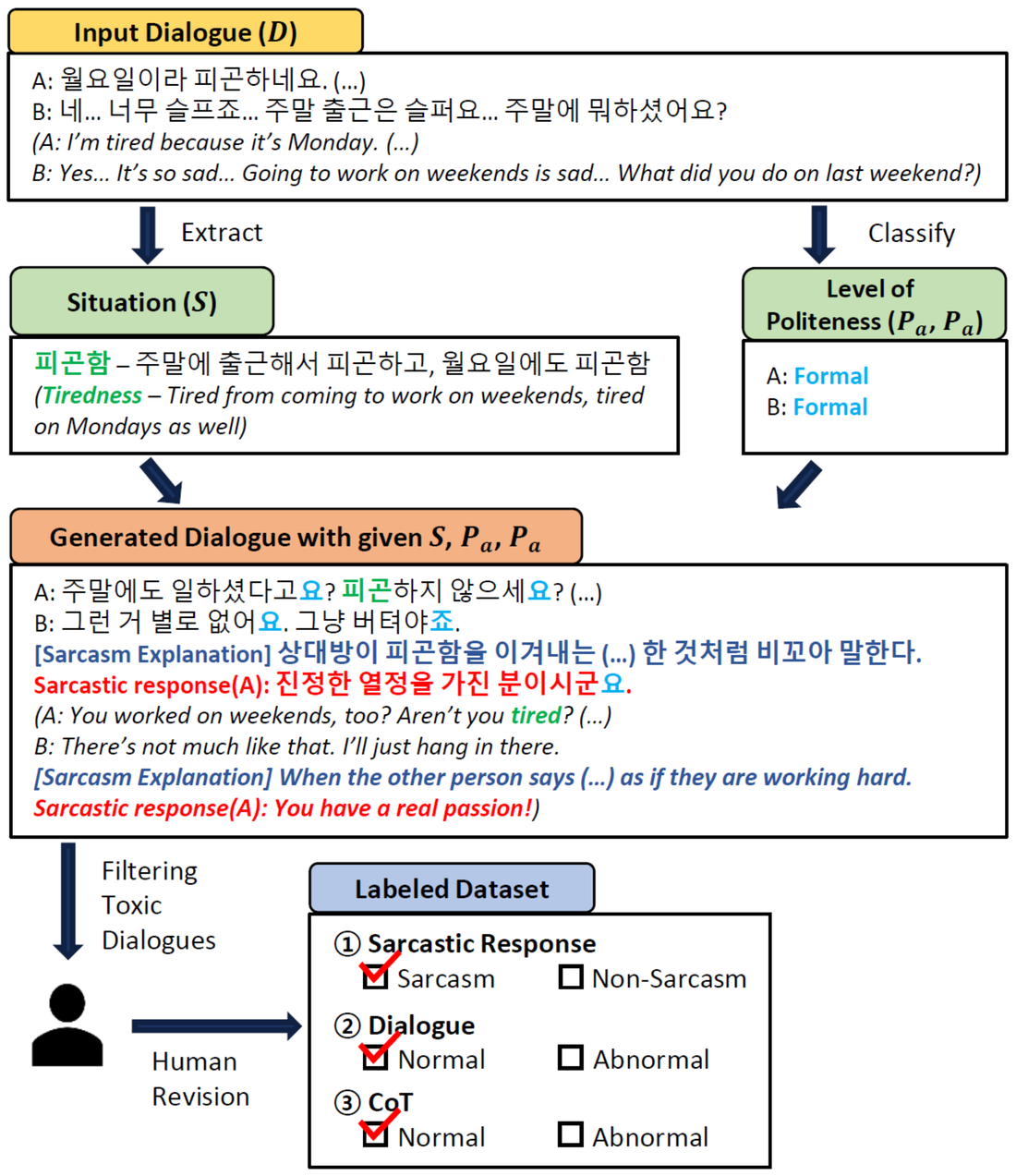
Motivation

- Non-English sarcasm datasets are not abundant. Sarcasm reflects cultural differences, so relying solely on translationese may not capture the linguistic nuances in sarcasm detection.
- There is only one available Korean sarcasm dataset but it lacks the necessary contexts.

Our Contributions

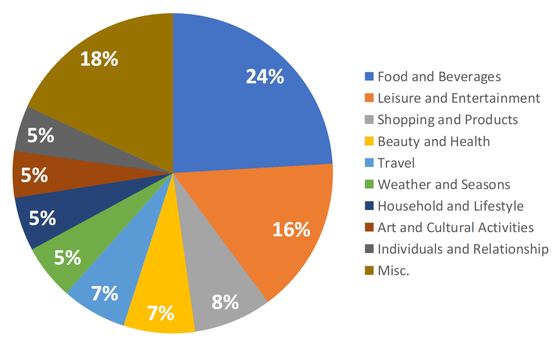
- We propose a comprehensive **dataset generation pipeline** for the context-aware sarcasm detection task using **both LLMs and human revision**.
- We introduce a new **large-scale Korean Context-aware Sarcasm detection dataset (KoCoSa)** through the proposed pipeline, which is composed of 12.8k daily dialogues.
- We provide a decent analysis of the Korean context-aware sarcasm detection task through this dataset, including the **strong baseline system** for the task.

Sarcasm Detection Dataset Construction



Dataset Statistics

Total Dialogues	12824
Sarcasm	7608(59.3%)
Non-Sarcasm	5216(40.7%)
Average Turns per Dialogue	4.3
Max Turns	10
Min Turns	2
Tokens per Dialogue	40.3
Tokens per Utterance	9.3
Tokens per Explanation	18.9



Experimental Results

Model	BA	M-F1	W-F1	Precision-S	Recall-S	Precision-N	Recall-N
Zero/Few-shot							
GPT-3.5(zero-shot)	53.5	43.4	40.6	71.2	20.5	40.2	86.5
GPT-3.5(4-shot)	51.8	42.1	39.4	66.4	20.0	39.2	83.5
GPT-3.5(8-shot)	49.4	32.3	27.2	57.7	6.0	37.8	**92.9**
GPT-4(zero-shot)	73.2	71.7	72.6	**83.3**	68.8	60.5	77.6
GPT-4(4-shot)	75.0	75.1	76.6	80.5	82.3	70.2	67.7
GPT-4(8-shot)	74.5	73.9	75.1	81.9	76.3	65.4	72.6
Fine-tuning							
KLUE-RoBERTa$_{base}$	74.0(±0.2)	74.1(±0.6)	74.7(±0.2)	71.5(±0.2)	**93.4**(±0.5)	**87.2**(±0.7)	54.7(±0.7)
KLUE-RoBERTa$_{large}$	74.9(±0.3)	75.1(±1.0)	75.5(±0.3)	74.6(±0.3)	85.0(±0.6)	80.0(±0.6)	64.8(±0.6)
Human Evaluation	**80.2**	**80.1**	**80.3**	83.0	80.5	77.1	79.9

- Fine-tuning models provide competitive scores compared to zero/few shot models despite the huge gap in model size.

Topic	Balanced Acc.		Weighted F1	
	KLUE	GPT	KLUE	GPT
Food and Beverages	71.9	73.2	71.7	73.9
Leisure and Entertainment	73.6	72.8	74.3	73.0
Individuals and Relationship	76.1	76.8	76.8	77.2
Beauty and Health	73.5	74.7	74.5	76.2

Context	Model	Human
Only Response	73.2	62.2
Last 1 Utterance + Response	73.2	-
Last 2 Utterance + Response	74.9	-
Last 3 Utterance + Response	75.8	-
Full Context	**76.0**	**80.2**

- We find that language models benefit from sufficient contextual information to enhance the accuracy of sarcasm detection.
- We demonstrate that there is not much difference in performance among topics.