Multimodal Sarcasm Explanation and Target Generation

by
Giridhar S

Under the supervision of
Dr. Md. Shad Akhtar

Submitted in partial fulfillment of the requirements for the degree of Master of Technology, CSE-AI

Department of Computer Science Engineering
Indraprastha Institute of Information Technology - Delhi
May, 2023

Certificate

This is to certify that the thesis titled *"Multimodal Sarcasm Explanation and Target Generation"* being submitted by **Giridhar S** to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

May,2023

Dr Md. Shad Akhtar

Department of Computer Science Engineering
Indraprastha Institute of Information Technology Delhi
New Delhi 110 020

Acknowledgements

I am deeply grateful to Dr. Md Shad Akhtar, my advisor, for granting me the incredible opportunity to embark on an independent research project and delve into the captivating realm of sarcasm understanding and comprehension. Throughout my thesis and academic journey, Dr. Akhtar's unwavering support and expert guidance have been invaluable, consistently pointing me in the right direction.

I would also like to extend my heartfelt appreciation to Dr. Tanmoy Chakraborty, whose honest insights, extensive knowledge, and encouragement have played a pivotal role in my growth. Dr. Chakraborty's guidance and insistence on exploring diverse domains have pushed me beyond my comfort zone, allowing me to broaden my perspective.

Furthermore, I owe a debt of gratitude to my family and friends for their constant moral and emotional support. Their unwavering presence has been a source of strength and motivation, making this endeavor possible

Abstract

Sarcasm is a nuanced and context-dependent communication that poses a significant challenge for natural language processing (NLP) systems. This study proposes a novel approach to improving sarcasm understanding by generating explanations and targets for sarcastic input using a multi-modal encoder-decoder transformer model. Our approach builds on an existing dataset called MORE by adding 2000 more instances and annotating the target of ridicule for all the sarcastic instances. The key idea behind our approach is to provide the model with more contextual information and common-sense knowledge to better understand sarcasm's complex and subtle nature. The model can give users a more accurate and adequate understanding of the underlying sarcasm by generating explanations and targets for sarcastic input. Our experimental results show that our proposed approach gives better results than the existing benchmark on the task of sarcasm explanation. Furthermore, our approach is highly interpretable, as the generated explanations and targets provide valuable insights into the model's decision-making process. Overall, our study demonstrates the effectiveness of our proposed approach for improving sarcasm understanding in NLP systems. Our approach has critical practical applications in sentiment analysis, opinion mining, and social media monitoring, where sarcasm understanding is crucial.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Definition

Sarcasm is a challenging sentiment to understand. The Oxford Dictionary defines sarcasm as "the use of irony to mock or convey contempt." It involves saying something opposite to the person's meaning, often intending to criticize or ridicule something or someone.

1.2 Motivation

Sarcasm is a complex form of sentiment to understand. There have been numerous studies on sarcasm detection, but just one study on sarcasm explanation. More than just detecting the sarcasm may be required for the users to understand why the given input is sarcastic. Hence, we improve the Multimodal Sarcasm Explanation (MUSE) task introduced in 'Desai, Poorav, Tanmoy Chakraborty, and Md Shad Akhtar—2022', where we generate explanations of the intended sarcasm for a given sarcastic input. Also, the target of sarcasm may not be limited to just the image and text. It can also be an action. Most studies focus on finding the target of sarcasm in either text or image. We propose a task where we generate the target of the sarcasm for a given sarcastic input. To this end, we annotate the dataset with the target of the ridicule in the sarcasm.

1.3 Background

Sarcasm refers to words that mean the opposite of what a person really wants to say, especially to insult someone, to show irritation, or to be funny. Sarcasm is a complex sentiment to detect. The early studies in sarcasm detection involved using only text data. In 'González-Ibáñez et al. (2011) [8]', the authors create a corpus for detecting sarcasm and use lexical and pragmatic features with machine learning to detect sarcasm in tweets. The authors also compare the scores obtained using machine learning with those obtained using human evaluation, showing that automatic classification works as well as human classification. But this study did not take into account the incongruity in the text. "Joshi et al. (2015) [11] " showed that incongruity in text helped detect whether a text was sarcastic. The incongruity is classified into implicit and explicit incongruity, where implicit refers to the sentences having opposite polarity to some words like "I loved this movie so much that I left in the middle of it," wherein "I left in the middle of it" has an opposite polarity to the word "love" and explicit referred to the words having opposite polarities like 'I love being ignored' where the word 'love' has an opposite polarity to the word 'ignored'. But this study did not capture the semantic and syntatic meaning of the words.

In 'Joshi, et al. 2016 [12]', the author then tried adding word-based embeddings such as word2vec, glove, etc. and saw an improvement in sarcasm detection. This show that semantic and syntatic meaning of the words help in understanding sarcasm. However, most works on sarcasm detection before 2019 were performed on textual data. In 'Castro et al. 2019 [2]', authors show that introducing multimodal cues improves sarcasm detection. The authors propose a novel dataset Multimodal Sarcasm Detection Dataset (MUStARD), enabling the development of multimodal approaches in sarcasm detection. In 'Sangwan et al.2021 [24]', the authors use deep learning to combine visual and textual modalities. It is a known fact that for sarcasm to work, the recipient should have common sense or knowledge of it. In 'Li, Jiangnan et al. 2021', the author integrates commonsense in sarcasm detection and shows its effectiveness on three datasets. These studies show that introducing multimodality and common sense help in understanding the sarcasm better.

In 'Desai et al. (2022) [7]', the author proposes a novel problem, 'Multimodal Sarcasm Explanation'. The task takes a multimodal (image and its caption) sarcastic post as input and aims to generate a natural language sentence to explain the intended irony in the sarcastic post. The author generates a novel multimodal sarcasm explanation dataset named MORE.

Previous studies on sarcasm target identication use only textual data. 'Wang et al. (2022)' [28] show that multimodality helps in sarcasm target identification.Studies related to sarcasm target identification either find the target in the text or image. But target need not be bounded to image and text.

1.3.1 Problem Definition

For a given multimodal sarcastic post $P = \langle I, T[t_1, t_2, ..., t_N] \rangle$, we aim to reveal the underlying sarcasm by generating a natural language explanation $E[e_1, e_2, ..., e_D]$, where $\forall t_i, e_j \in$ VocabEnglish.

We also propose a novel task of sarcasm target generation. For a given multimodal sarcastic post $P = \langle I, T[t_1, t_2, ..., t_N] \rangle$, we aim to generate the target of ridicule in the sarcastic input $R[r_1, r_2, ..., r_D]$, where $\forall t_i, r_j \in$ VocabEnglish.

.

1.4 Contribution

Our contributions are as follows:

- Sarcasm always has a target of ridicule. In this paper, we extend the MORE dataset by providing a target of the sarcastic input.
- We also increase the number of instances in the dataset by 2000 instances. Thus, the total number of instances in the dataset is 5510, consisting of image, caption, explanation, and target.
- We introduce a novel multimodal transformer encoder-decoder model that generates explanations and targets.
- Using the model, we show that explanation improves the quality of the target, and the target improves the quality of the explanation.

Chapter 2

Related Work

2.1 Sarcasm Detection:

Sarcasm detection tasks were initially performed using unimodal approaches such as text.In "Davidov et al. (2010) [6]", the authors propose a method for sarcasm detection based on lexical, syntactic, and semantic features. In "Ptáček, Tomáš, et al. (2014) [22]", the authors propose a method for sarcasm detection based on sentiment analysis and pattern matching. In 'Joshi et al. (2015) [11]', the authors identify text incongruity as a critical sign of sarcasm. 'Bouazizi and Otsuki Ohtsuki (2016) [1]' use sarcasm-related features, including sentiment, punctuation, semantic, and pattern-related features, to detect sarcasm. However, these studies primarily focus on texual data for sarcasm detection.In "Poria et al. (2016) [21]", the authors propose a method for sarcasm detection using a deep learning model that combines textual, visual and acoustic features. 'Castro, et al. (2019) [3]' propose a multimodal approach by providing a Multimodal Sarcasm Detection Dataset (MUSTARD). 'Sangwan et al. (2020) [24]' use textual and visual modalities to detect sarcasm by providing more clues. 'Y. Wu et al. (2021) [29]' proposed an incongruity-aware attention network (IWAN), which detects sarcasm by focusing on the word-level incongruity between modalities via a scoring mechanism.

2.2 Sarcasm Detection using Commonsense Knowledge:

'Li, Jiangnan, et al [16]' propose a novel deep-learning architecture for sarcasm detection by integrating commonsense knowledge. In this paper, the authors used a pre-trained

COMET model to generate relevant commonsense knowledge. 'Chowdhury, et al. 2021 [4]' use COMET and pretrained embeddings along with graph convolution network to show that incorporating commonsense knowledge's improves sarcasm detection.

2.3 Natural Language Generation:

Recents studies have used multimodality for natural language generation 'Hendricks et al. (2016) [9]' used a fine-grained classifier to predict a label and then concatenated the label with the visual features to explain the predicted label. 'Costa, Felipe, et al. (2018) [5]' generated explanations using user reviews. Elliott et al. (2013) [18] introduces a model that generates natural language descriptions based on visual perception, incorporating image features to enhance the language generation process. Karpathy and Fei-Fei (2015) [13] proposes an influential model that generates image descriptions by combining convolutional neural networks (CNN) for visual understanding and recurrent neural networks (RNN) for language generation. Xu et al. (2015) [30] presents an attention-based image captioning model that attends to different parts of the image while generating captions, resulting in more detailed and accurate descriptions. Das et al. (2020) [19] introduces an approach that incorporates image-based context into dialogue generation, enhancing the generation of questions and responses in a multimodal conversational setting.

2.4 Sarcasm Target Identification:

'Joshi et al. (2018) [10]' introduced a novel dataset where the sarcasm target of the tweet or book snippet is annotated. The authors then used rule-based and statistical-based extractors to identify the sarcasm target. 'Wang et al. (2022)' [28] introduced multimodality to sarcasm target identification and showed that visual cues help find the target of sarcasm in the text and image with reasonable accuracy.

2.5 Sarcasm Explanation:

'Kumar et al. (2022) [14]' studied the discourse structure of sarcastic conversations and proposed a novel task - Sarcasm Explanation in Dialogue (SED). The task aims to generate natural language explanations of satirical conversations. The authors also curated a new WITS dataset to support the task. They proposed MAF (Modality Aware Fusion), a multimodal context-aware attention and global information fusion

module to capture multimodality and use it to benchmark WITS. 'Desai et al. (2022) [7]' proposed a novel problem - Multimodal Sarcasm Explanation (MuSE) - where given a multimodal sarcastic post containing an image and a caption, the authors aimed to generate a natural language explanation to reveal the intended sarcasm. They also developed a novel MORE dataset containing 3510 sarcastic multimodal posts.

Chapter 3

Dataset

3.1 MORE Dataset

We extend the Multimodal sarcasm dataset (MORE) proposed by "Desai et al. (2022) [7]". The dataset initially consisted of 3510 instances. Each instance consisted of a tuple (image, caption, explanation). To train our model on a large dataset, we extend the dataset by annotating 2001 more instances. Thus the extended dataset has 5511 instances and is split into train, val, and test in the ratio 85:5:10. We also annotate the target of ridicule for all the instances in the dataset. The annotation guidelines from Desai et al. (2022) [7] is adopted in this study.

3.2 Target of Sarcasm

Sarcasm is characterized by a target of ridicule, which can be an object, action, or other entity. In the case of multimodal sarcastic inputs, the target may be present in the accompanying image or text, or it may not be present in either. To address this challenge, we propose a task in which the target of ridicule in a given sarcastic input is generated rather than identified in the image or text. We have annotated 5511 target instances in the MORE dataset to facilitate this task. Each instance in the dataset comprises an image, a caption, an explanation, and the corresponding target entity, forming a tuple.

3.3 Dataset Statistics

The original and extended datasets are provided in Table 3.1 and Table 3.2, respectively. The original dataset is split into train(2983), val(175), and test(352) in the ratio 85:5:10.

split	# of posts	Caption		Explanation	
		Avg.length	$\|V\|$	Avg.length	$\|V\|$
Train	4692	15.31	16459	11.99	9012
Val	266	14.87	1807	12.02	1239
Test	553	15.13	3374	11.87	2148
Total	5511	15.10	21640	11.96	12399

Table 3.1: Original Dataset Statistics

split	# of posts	Caption		Explanation		Target	
		Avg.length	$\|V\|$	Avg.length	$\|V\|$	Avg.length	$\|V\|$
Train	4692	15.31	16459	11.99	9012	4.48	5108
Val	266	14.87	1807	12.02	1239	4.62	634
Test	553	15.13	3374	11.87	2148	4.73	1223
Total	5511	15.10	21640	11.96	12399	4.61	6965

Table 3.2: Extended Dataset Statistics

The extended dataset is also split using the ratio 85:5:10 into train (4692), val(266) and test(5511). The length of the explanation is not big, showing that the sarcastic input can be explained concisely.

3.4 Annotation

We adopt the annotation guidlines followed for the MORE dataset. The following are the annotation guidlines:

- Non-Sarcastic posts are discarded.
- Sarcastic posts explicitly mentioning the word sarcasm are discarded.
- All entities, including images, captions, hashtags, emojis, Etc., are considered to understand the underlying sarcasm and generate the appropriate explanation and target.
- In the case of sarcasm being explained in multiple ways, a shorter and more precise explanation is preferred.
- Target need not be bounded only to text or image.

Caption: awesome ! look @ all the taxis waiting to take arrival passengers home !

Explanation: there's no taxi waiting to take arrival passengers home.

Target: no taxis for the passengers

Figure 3.1: Example of a sarcastic instance in the dataset

Chapter 4

Proposed Methodology

We propose a model architecture that generates explanations and targets from input images and captions. Our model utilizes the CLIP transformer (Radford et al. (2021) [23]) introduced to generate image embeddings and the BART encoder to generate caption and target embeddings. Additionally, we initialize the explanations and targets to "none" and generate their embeddings using the BART encoder. Using the BART encoder, we join the caption and target using the separator token ¡sep¿ and convert the concatenated sequence to embeddings. Similarly, we join the caption and explanation using ¡sep¿ and convert the concatenated sequence to embeddings using the BART encoder. We use word2vec embeddings trained on the ConceptNet graph (Speer et al. 2017 [26]) to introduce knowledge into the model.

Our model architecture comprises two modules: the explanation and the target modules. We feed the image embeddings, target embeddings, and the joint caption and target embeddings into the explanation module. In contrast, we feed the image embeddings, explanation embeddings, and joint caption and explanation embeddings into the target module. We fuse these embeddings using the multi-head attention module. The output of the explanation module's explanation is fed back to the target module, image embeddings, and caption embeddings, and vice versa.

4.1 Image Encoder

We use CLIP (Contrastive Language-Image Pretraining) model [23] as image encoder. The input image is passed to the CLIP model and its last hidden state is considered as the image representation ($I \in R^{q \times d}$). Here q is the sequence length and d is the hidden size of the layer.

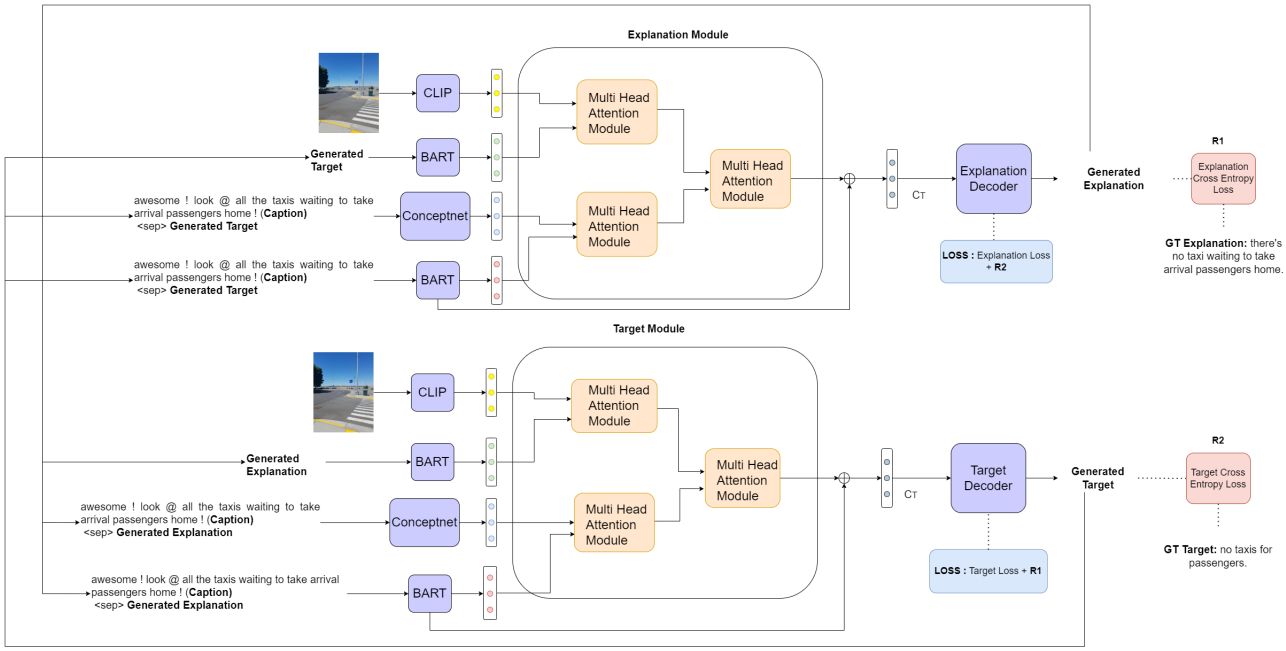


Figure 4.1: Architecture of TExMoret

4.2 Text Encoder

We use BART Encoder as the text encoder. The input caption, generated target and generated explanation are passed through BART Encoder to get their respective representations ($T \in R^{p \times t}$). Here p represents the number of tokens in the text input and t represents the dimension of each token representation.

4.3 Conceptnet

ConceptNet is a knowledge graph that represents general knowledge and common sense about the world in a machine-readable format. It is designed to support natural language understanding and reasoning tasks by providing a network of concepts and their relationships.We provide commonsense to the model by converting each token in the text input into the Conceptnet based word2vec embeddings.

4.4 Explanation Module

The explanation module consists of three components. The first component is the fusion of image and target embeddings using the multi-head attention module. The second component fuses the BART and ConceptNet-based word2vec embeddings of the concatenated caption and target sequence. The third component takes the output of the first and second components and generates a fused embedding using multi-head attention. The output embedding is concatenated with the BART embedding of the concatenated caption and target sequence and provided to the BART decoder to generate explanations.

4.5 Target Module

Similarly, the target module has three components. The first component is the fusion of image and explanation embeddings using the multi-head attention module. The second component fuses the BART and ConceptNet-based word2vec embeddings of the concatenated caption and explanation sequence. The third component generates a fused embedding using multi-head attention, taking the output of the first and second components as input. The output embedding is concatenated with the BART embedding of the concatenated caption and explanation sequence and provided to the BART decoder to generate targets.

4.6 Multi Head Attention Module

We employ the multi-head attention module that Vaswani et al. (2017) introduced, where the Query and Key vectors' scaled dot product are performed and multiplied with the Value vector. We apply this attention over multiple heads, where the image embeddings are the Value and Key vectors, and the target and caption embeddings are the Query vectors.

$$Attention(Q, K, V) = softmax(QK^T/d^k)V \tag{4.1}$$

4.7 Fully Connected Layer

We use a 2-layered fully connected feed forward network with a ReLU activation. Both the layers used are of size 2048.

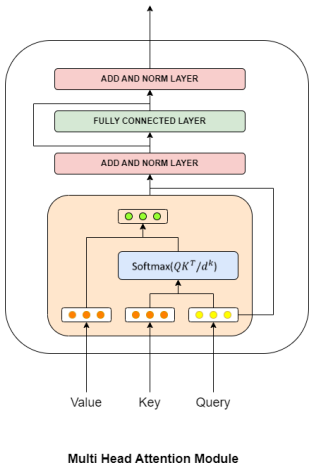


Figure 4.2: Multi Head Attention Module

4.8 Explanation Decoder

In our approach, we employ the BART Decoder to generate explanations and targets. To create the final encoder representation, we concatenate the output from the Explanation module (d) with the caption representation. This encoder representation (D_{T}) is then provided to the BART Decoder, along with the decoder inputs. Through an autoregressive process, the BART Decoder generates explanations and targets.

4.9 Intuition of the Proposed Methodology

We adopt the intuition that providing a target helps the model generate proper explanations, and providing an explanation helps the model generate proper targets for the sarcastic input. To implement this intuition, we return the generated explanations and the target to the model so that the model learns to generate better explanations and targets.

Chapter 5

Results

5.1 Experimental Setup

We conducted experiments on the MORE dataset and partitioned it into 85% for training (4692 samples), 5% for validation (266 samples), and 10% for testing (553 samples). To evaluate the generated explanations, we utilized a range of standard metrics, including BLEU (B1, B2, B3, and B4), ROUGE (R1, R2, and RL), METEOR, BERTScore (Zhang et al., 2020), and SentBERT. SentBERT is a BERT-based model that computes the cosine similarity between the reference and generated explanations in the Sentence-BERT embedding space, which measures their semantic closeness.

5.2 Hyperparameters:

We employ BART (Lewis et al. 2020 [15]) tokenizer with maximum token length as 256. We employ CLIP (Radford et al. (2021) [23]) as image encoder and Conceptnet based word2vec embeddings. We use AdamW (Loshchilov and Hutter 2017 [17]) optimizer with learning rate of 1e-5 for both the cross-modal encoders and 3e-4 for the LM head of decoder. We train TExMoret for 40 epochs with batch size = 8. During training, the cross-entropy loss is monitored over the validation set with image encoder in a frozen state.

5.3 Baselines:

We adopt the same baselines used in 'Desai et al. (2022) [7]' and train them on the extended dataset.

5.3.1 Text Based:

We employ transformer [27] and pointer-generator networks [25] as text based baselines for generating explanations and targets. Both these baselines have used in summarization tasks and only use text modality as input.

5.3.2 Text and Image Based:

We adopt the multimodal transformer introduced by Yao and Wan, 2020 [31] for the task of multimodal machine translation as one of the baseline. We adopt the ExMore model proposed by "Desai et al. (2022) [7]" as the baseline for our sarcasm explanation task. To generate targets for the task of target generation, we employ the architecture of ExMore to generate targets and use it as our baseline model.

5.4 Evaluation Metrics

We employ BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-1, ROUGE-2, ROUGE-L, METEOR, BertScore and cosine similarity between the sentence level embeddings of the ground truth and generated output. The cosine similarity shows how contextually similar the generated output and the ground truth are.

5.5 Loss:

We use the cross-entropy loss function to measure the difference between the predicted and actual probability distributions. The explanation loss is the cross-entropy loss between the predicted logits of the explanation module and the target IDs of the ground truth explanation. Similarly, the target loss is the cross-entropy loss between the predicted logits of the target module and the ground truth target of ridicule target IDs. To reinforce the explanation loss of the explanation module, we add the target loss, and to reinforce the target loss of the target module, we add the explanation loss.

$$E_{\mathrm{L}} = E_{\mathrm{L}} + T_{\mathrm{L}}\ (R_2)$$

$$T_{\mathrm{L}} = T_{\mathrm{L}} + E_{\mathrm{L}}\ (R_1)$$

Where E_{L} = Explanation loss and T_{L} = Target Loss

R1 and R2 are the reinforced target and explanation losses respectively.

Model	BLEU				Rogue			METEOR	BERT-Score			Sent-BERT (Cosine)
	B1	B2	B3	B4	R1	R2	RL		Pre	Rec	F1	
Transformer	14.31	6.4	3.9	2.3	21.22	8.37	20.12	17.71	84.7	85.2	84.94	53.1
Pointer Generator Networks	21.46	8.96	3.9	2.3	22.14	8.48	20.9	22.79	85.51	86.35	85.92	52.8
M-Transf	17.53	8.36	4.1	2.67	26.29	8.94	23.71	23.12	87.81	87.1	87.42	54.72
ExMore	23.27	14.37	8.34	5.06	32.05	15.94	29.96	28.62	89.1	88.0	88.5	59.93
TExMoret	27.75	18.03	10.65	6.75	37.44	19.67	34.91	33.79	89.5	88.8	89.1	62.79

Table 5.1: Evaluation metrics for ExMore and TExMoret models for sarcasm explanation

Model	BLEU				Rogue			METEOR	BERT-Score			Sent-BERT (Cosine)
	B1	B2	B3	B4	R1	R2	RL		Pre	Rec	F1	
Transformer	4.49	0.93	0.11	0.031	12.13	1.99	9.81	11.23	81.70	81.11	81.4	41.87
Pointer Generator Networks	9.58	1.79	0.69	0.088	12.17	2.06	12.07	14.99	82.29	81.9	82.09	40.39
M-Transf	7.46	1.93	0.74	0.093	14.57	2.34	14.19	13.68	83.35	82.97	83.16	42.66
ExMore	11.27	4.6	1.8	0.31	19.69	7.74	19.52	15.9	85.3	84.6	84.94	46.22
TExMoret	14.21	6.08	2.84	0.94	22.65	9.49	22.20	16.98	86.4	85.6	86.0	48.79

Table 5.2: Evaluation metrics for ExMore and TExMoret models for sarcasm target generation

5.6 Experimental Results

Table 5.1 shows the comparison of our model (TExMoret) with the baselines for sarcasm explanation. From Table 5.1 we can see that TExMoret outperforms our best baseline model ExMore across all the evaluation metrics. We obtain BLEU scores of 27.75(+4.48), 18.03 (+3.66), 10.65 (+2.31), 6.75 (+1.69) for B1, B2, B3 and B4 respectively. Similarly we obtain Rouge scores of 37.44 (+5.39), 19.67 (+3.73), 34.91 (+4.95) for R1, R2 and RL respectively. Similarly we see improved scores in metrics such as METEOR and BERT-Score. We obtain a cosine similarity of 62.79 (+2.86) using SentBERT showing that explanation generated by our model has better context.

Similarly from Table 5.2 shows the comparison of our model (TExMoret) with the

Model	POS tags			
	Noun	Verb	Adjective	Adverb
Ref Count	3.46	1.39	1.14	0.57
ExMore Gen count	3.02	0.87	0.95	0.36
TExMoret Gen count	3.51	1.08	0.85	0.31

Table 5.3: Counts of POS Tags for ExMore and TExMoret generated explanations

Model	BLEU				Rogue			METEOR	BERT-Score			Sent-BERT (Cosine)
	B1	B2	B3	B4	R1	R2	RL		Pre	Rec	F1	
Explanation	24.46	15.77	9.64	6.22	34.88	17.85	32.67	31.16	89.3	88.5	88.89	61.33
Target	13.7	5.4	2.2	0.6	20.25	8.32	21.12	16.3	85.9	85.03	85.54	47.94

Table 5.4: Evaluation metrics for alternate version of TExMoret models for explanations and targets generation

baselines. It can be seen that TExMoret outperforms our best baseline model ExMore across all the evaluation metrics. We obtain BLEU scores of 14.21(+2.94), 6.08 (+1.48), 2.84 (+1.04), 0.94 (+0.63) for B1, B2, B3 and B4 respectively. Similarly we obtain Rouge scores of 22.65 (+2.96), 9.49 (+2.05), 22.20 (+2.68) for R1, R2 and RL respectively. Similarly we see improved scores in metrics such as METEOR and BERT-Score. We obtain a cosine similarity of 48.79 (+2.57) using SentBERT showing that explanation generated by our model also has better context.

5.7 Ablation study

In our experimentation, we have developed an alternative version of TExMoret that does not incorporate targets and explanations as input. Our evaluation, as shown in Table 5.4, indicates that this new version of TExMoret exhibits lower performance when compared to our original model. This observation underscores the importance of incorporating a suitable target for enhancing sarcasm explanation and providing an appropriate explanation for improving the sarcasm target generation.

5.8 Result Analysis

In this section, we conduct a linguistic analysis of the generated explanations by comparing them based on four content parts-of-speech (POS) tags: noun, verb, adjective, and adverb. Since POS tags convey the semantics of sentences, comparing their frequency in the generated (Gen) and reference (Ref) explanations can offer valuable insights into their semantic context. The average counts of the POS tags over the test set are presented in Table 5.3. The findings indicate that TExMoret performs similarly to the reference counts in two of the four POS tags, whereas EXMORE performs worse than the reference counts in all four POS tags.

Model	Adequacy (Explanation)	Fluency(Explanation)	Adequacy(Target)
EXMORE	0.58	4.12	0.43
TExMoret	0.67	4.42	0.54

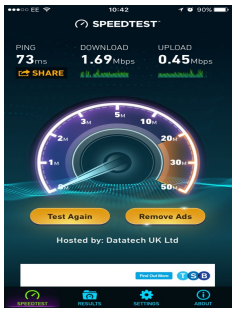
Table 5.5: : Adequacy and Fluency of explanations by ExMore and TExMoret

5.9 Human Evaluation

We conduct human evaluations to assess the quality of the generated explanations and targets and follow evaluation used in 'Desai et al. (2022) [7]'. We sample 50 instances from the test dataset and provide them to 25 evaluators with the generated explanation and target of EXMORE and TExMoret. The evaluators rate the output usng two metrics - adequacy and fluency. Adequacy represents the appropriateness of the generated output whereas fluency represents the consistency of the generated output in English. For adequacy the human evaluators are given for choices - justify, weakly justify, somewhat related to input (SRI) and not related to input (NRI). Justify represents the high similarity between the generated explanation or target with their respective ground truth. In case of explanation, weak justify represents explanation which give semantic incongruence but does not explain the sarcastic nature whereas in case of target it represents target is good but not perfect. SRI represents the explanation and target with words strongly related to input where NRI represents no relation to the input.

We assign a score of 1.0 to justify, 0.66 to weakly justify, 0.33 to SRI, and 0.0 to NRI samples. Evaluators rate fluency of the generated explanation on the discrete scale of [1,5]. We don't rate fluency for the generated target as the number of tokens in the target are very less. Table 5.5 presents the summary of human evaluation in form of adequacy and fluency scores. From Table 5.5, we observe that the evaluators showed more confidence in the explanation of TExMoret than ExMore for both adequacy (0.69 vs 0.37) and fluency (0.67 vs 0.58) metrics. Similarly, the evaluators showed more confidence in target of TExMoret than ExMore.

To compute the adequacy rating distribution, we adopt the majority-voting approach across evaluators to select the adequacy class. The results are shown in Table 5.6. We observe a significant percentage of samples fall under the justify or weakly justify categories for TExMoret. In contrast, most of the samples belong to the SRI and NRI categories for EXMORE. It further strengthens our claim that TExMoret yields better explanation and target than all baselines.



Caption: <user>so pleased with download speed right now ! timetoswitch

GT Explanation: the author is pissed at <user>for such low download speed.

Pred Explanation: the author is pissed at <user>for such terrible speed

GT Target: <user>low download speed

Pred Target: slow internet speed

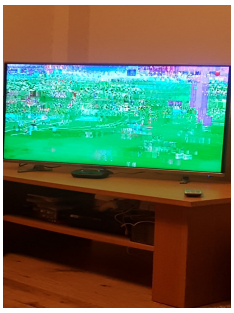
Caption: i would like to personally thank <user>for being the best company out there . this makes 17/18 bad experiences . solid record .

GT Explanation: the author is pissed at<user>for such bad packaging.

Pred Explanation: the author is pissed at <user>for being such terrible customer service.

GT Target: <user>'s bad packaging

Pred Target: <user>'s bad customer service



Caption: really enjoying the atmfcb game tonight <user>

GT Explanation: the author can't enjoy the atmfcb game tonight because of such disturbance.

Pred Explanation: the author is pissed at <user>for such a game.

GT Target: disturbance in the channel

Pred Target: bad quality of the game

Caption: nice parking jobs , guys.

Ground Truth: these are bad parking jobs, the cars are out of the slots.

Explanation: the author is pissed at ¡user¿.

GT Target: bad parking by the people

Pred Target: bad parking

Figure 5.1: Examples of Adequacy Ratings (Justify, Weakly Justify, Relatedto Input, Not Related to Input) for explanations from left to right respectively.

Model	Adequacy Rating			
	Justify	Weakly Justify	Strongly related to Input	Nout related to Input
ExMore	50%	10%	30%	10%
TExMoret	60%	10%	25%	5%

Table 5.6: Adequacy rating distribution

Chapter 6

Conclusion

Based on our experimental results and analyses, our proposed model TExMoret outperforms the baseline models and achieves state-of-the-art performance on the sarcasm target generation and explanation tasks. The performance improvement can be attributed to incorporating targets, explanations, and common sense, in our models. Overall, our work contributes to the advancement of sarcasm understanding in natural language processing and provides insights into the importance of explanation and targets in sarcasm analysis. Future work can focus on improving the semantic-level analysis of generated explanations and exploring the potential of incorporating other forms of contextual information in sarcasm interpretation.

Bibliography

[1] Mondher Bouazizi and Tomoaki Otsuki Ohtsuki. A pattern-based approach for sarcasm detection on twitter. *IEEE Access*, 4:5477–5488, 2016.

[2] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy, July 2019. Association for Computational Linguistics.

[3] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815*, 2019.

[4] Somnath Basu Roy Chowdhury and Snigdha Chaturvedi. Does commonsense help in detecting sarcasm? *arXiv preprint arXiv:2109.08588*, 2021.

[5] Felipe Costa, Sixun Ouyang, Peter Dolog, and Aonghus Lawlor. Automatic generation of natural language explanations. In *Proceedings of the 23rd international conference on intelligent user interfaces companion*, pages 1–2, 2018.

[6] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Semi-supervised recognition of sarcasm in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 107–116, 2010.

[7] Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10563–10571, 2022.

[8] Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Annual Meeting*

of the Association for Computational Linguistics: Human Language Technologies, pages 581–586, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[9] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 3–19. Springer, 2016.

[10] Aditya Joshi, Pranav Goel, Pushpak Bhattacharyya, and Mark Carman. Sarcasm target identification: Dataset and an introductory approach. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[11] Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762, Beijing, China, July 2015. Association for Computational Linguistics.

[12] Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark Carman. Are word embedding-based features useful for sarcasm detection? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1006–1011, Austin, Texas, November 2016. Association for Computational Linguistics.

[13] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[14] Shivani Kumar, Atharva Kulkarni, Md Shad Akhtar, and Tanmoy Chakraborty. When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues. *arXiv preprint arXiv:2203.06419*, 2022.

[15] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[16] Jiangnan Li, Hongliang Pan, Zheng Lin, Peng Fu, and Weiping Wang. Sarcasm detection with commonsense knowledge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3192–3201, 2021.

[17] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 11 2017.

[18] Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. A joint model of language and perception for grounded attribute learning. *arXiv preprint arXiv:1206.6423*, 2012.

[19] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P Spithourakis, and Lucy Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*, 2017.

[20] Anupiya Nugaliyadde, Kok Wai Wong, Ferdous Sohel, and Hong Xie. Enhancing semantic word representations by embedding deep word relationships. In *Proceedings of the 2019 11th International Conference on Computer and Automation Engineering*, ICCAE 2019, page 82–87, New York, NY, USA, 2019. Association for Computing Machinery.

[21] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59, 2016.

[22] Tomáš Ptáček, Ivan Habernal, and Jun Hong. Sarcasm detection on czech and english twitter. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 213–223, 2014.

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[24] Suyash Sangwan, Md Shad Akhtar, Pranati Behera, and Asif Ekbal. I didn't mean what i wrote! exploring multimodality for sarcasm detection. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[25] Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017.

[26] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[28] Jiquan Wang, Lin Sun, Yi Liu, Meizhi Shao, and Zengwei Zheng. Multimodal sarcasm target identification in tweets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8164–8175, 2022.

[29] Yang Wu, Yanyan Zhao, Xin Lu, Bing Qin, Yin Wu, Jian Sheng, and Jinlong Li. Modeling incongruity between modalities for multimodal sarcasm detection. *IEEE MultiMedia*, 28(2):86–95, 2021.

[30] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[31] Shaowei Yao and Xiaojun Wan. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4346–4350, 2020.