Sarcasm Explanation In Multimodal Dialogues

by
Shashwat Vaibhav

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
December, 2022

Certificate

This is to certify that the thesis titled *"Sarcasm Explanation In Multimodal Dialogues"* being submitted by **Shashwat Vaibhav** 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.

December,2022

Dr Md. Shad Akhtar

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

Acknowledgements

I am grateful to my advisor Dr.Md Shad Akhtar for providing me this opportunity of carrying out an independent research project and to allow me to explore the topic of sarcasm understanding and comprehension.My advisor has provided support and guided me in the right direction throughout the duration of my thesis and academics.
I would also like to extend my gratitude towards Dr. Tanmoy Chakraborty for providing me with honest and knowledgeable insights, guidance and platform and for pushing me to explore variety of domains so that I could get out of my comfort zone.
I also thank my mentor Shivani Kumar who has provided me a critical support during the challenging intervals of my research work.
This work would not have been possible without the constant moral and emotional support from my family and friends.

Abstract

Sarcasm is a means to convey ridicule or contempt. There has been a plethora of work in sentiment analysis, of which sarcasm is one of the most challenging tasks due to the incongruity between the surface level and the intended meaning of the sarcastic remark. In a dialogue setting, comprehending an ironic utterance is challenging, especially when the context is unclear. The classical studies primarily dealt with sarcasm detection tasks which considered the textual modality as the prime one. These works did well in sarcasm detection but failed to provide any explanation behind the elicited sarcasm, resulting from the lack of understanding and comprehension of a sarcastic utterance. The hidden semantic meaning of a sarcastic utterance is difficult to grasp without complete contextual clues, such as acoustic and visual signals. To this end, we explore the task of multimodal sarcasm explanation in dialogues, which deals with generating a natural language explanation for any given sarcastic instance. In this work, we are proposing the Explanans (What to explain) and Explanandum (what does the explaining) and follow various psychological theories to explain the satirical discourse. We show quantitative and qualitative analysis of the proposed model and discuss a possible future venture which can involve the incorporation of cognitive features as well.

Contents

List of Figures

Chapter 1

Introduction

The word **sarcasm** has its origin in the Greek noun called ***sarkasmos*** which in its literal meaning implies *"a remark which is sneering and hurtful"*. Sarcasm is a means to ridicule, hurt and criticize someone.

The use of figurative language has been prevalent in our lives for ages. We use it to convey meanings beyond the stated fact. The scenario is more complex when there is an incongruity between the hidden meaning and the literal one. Classic examples of such cases are the presence of irony, humor, or satire. Sarcasm also belongs to this set.

The field of Natural Language Processing(***NLP***) deals with how machines can understand languages that we humans can use in our day-to-day lives. One such task is to find the sentiment behind a situation. Sarcasm is always intended to be negative. For example, a student arrives late to class. The professor says, *Wow! You are so early!*. Here the intended or the hidden meaning behind the utterance is the opposite of what had been said by the professor. It was meant to ridicule the student. Thus, the sentiment was negative.

The complexity associated with sarcasm is that the sentiment related to a sarcastic utterance is always misunderstood to be positive. For example, a person writing a review regarding a mobile phone writes a review, *"The battery life is sooo LONGGGG!"* despite the battery being of abysmal quality. If a typical machine learning pipeline is used to classify, it will label it as a sentence with a positive sentiment.

Sarcasm is always confused with humor and irony. It can be humorous and depict an ironic scenario but the associated sentiment is always negative. The question does arise of how to find whether a statement or a said utterance is sarcastic or not. The answer is not yet simple. Sarcasm is not always *explicit*. For example in the sentence, *"I love being excluded!"*. There are two sentiment-bearing words, ***love*** and ***excluded*** with positive and negative sentiments, respectively. The presence of opposite sentiment

words makes it a case of explicit sarcasm. This idea has been implemented as a rule-based approach in many of the earlier tasks in NLP for sarcasm detection. But this does not always work in cases where sarcasm is *implicit*. For example in the sentence, *"I love this paper so much that I will use it to play paper-throw."* In this case, the earlier paradigm will not work as there are no opposite sentiment-bearing words.

To overcome this, we need to ask if there are any markers of sarcasm. The presence of repetitions as in *"LONGGGGG"*, exclamation marks, hashtags, etc., are some of them, which are primarily used in rule-driven machine learning and deep learning pipeline. But these rule driven ways face challenges with the preprocessing step, where the semantics gets lost in between.

The rule-based paradigm suffers from a bottleneck of lack of context. For example, someone says *"I love this weather."* The statement can only be sarcastic if we know the proper context. By context, it means that if we know that the person hates the rainy season and the weather is indeed rainy, then only we can say that the person was sarcastic. Sarcasm is inherently context-dependent. Context gives the freedom to express something beyond words, beyond the obvious.

The research in NLP has dealt chiefly with text data. The problem with text is you can be descriptive and elaborative in your writing as much as you want, but the information conveyed by the text can not always be impactful and significant. Let's say an author has to write a scene where a person has to sarcastically say *"oh really!"*. The author can write it as "She said *"oh really!"*, with rolling her eyes". But can we be that descriptive all the time?

The answer lies in the use of different modalities. Different modes of information play different roles while conveying the overall context. The *acoustic* modality can help us to know the current mood and emotional state of the speaker depending on the intonation and modulation of their voice. The visual modality, which could be text or video, may help us to gauge the incongruity using visual cues such as facial expression, weather, place, etc. Considering multiple modalities helps us to know the unknown and fathom the unfathomable.

Multimodality has been explored in the area of NLP extensively in this era because of the promising results it has given. It helps us to reveal the incongruities that might be present at the modality level. For example, a person posts on Instagram, an image with the rainy weather depicting inundation, with a caption *"I love this weather"*. The inter-modality incongruence helps us to detect that the person was being sarcastic.

Sarcasm detection has been done in the research field for various needs. The machine learning models and gradual rise and transition to better-performing deep learning models are really helping in detecting sarcasm with improved accuracy. But most of these models are learning this classification task without actually inferring the rationale

behind something being sarcastic or non-sarcastic. The model might be giving the right results but the result itself could be based on incorrect reasons. This has led to the task of explanation of why something is sarcastic.

The task of sarcasm explanation generation is a sequence-to-sequence task that extracts the semantics from the data and generates the meaning behind the utterance. It involves Natural Language Understanding(***NLU***). From understanding the semantics, it has to infer the meaning beyond what has been said and then generate the explanation.

Dialog conversations play a crucial role in today's world of NLP. It is an immensely significant part of automated conversation systems such as chatbots. To generate a coherent response, the task has to go beyond the general sequence-to-sequence task. Sarcasm in dialog conversation is difficult as several caveats have to be considered. Sarcasm in a conversation could be *Intended* or *Perceived*. A person making an utterance, i.e., a source of sarcasm in a dialog might not have the intention of being sarcastic, but another participant in the same conversation might perceive it as sarcasm. There can be cases when the target of sarcasm might fail to perceive the mockery or ridicule. Hence to understand sarcasm in such cases and thereby generate an appropriate explanation behind the hidden meaning becomes really difficult.

A general theory of finding the hidden meaning behind an utterance is called ***Literal First Theory***. It is also known as ***standard Pragmatic Model***. It states that, given the presence of incongruity between the literal and the hidden meaning, one first understands the literal meaning and then finds if any contextual incongruity is present, then only they are able to comprehend the sarcasm. In one of the previous examples, a person has to first understand the literal meaning that the *weather is nice*. As the presence of inundation leads to semantic incongruity, the sarcasm behind that sentence is understood.

For a successful conversation to take place, there are certain psycho-linguistic theories that help us decode the process of such communication. Herbert Paul **Grice**, one of the pioneers of linguistic psychology, gave his famous co-operative principle. The co-operative principle states that during a conversation, there must some common ground or common knowledge on which successful communication could be based. During the process, four ***maxims of conversation*** must be followed, which are as follows:

1. **Maxim of Quality** (Be truthful): For communication to be successful, one should not lie.

2. **Maxim of Quantity** (Be specific): One should not exaggerate or give too less information.

3. **Maxim of Relevance** (Be relevant): One should be relevant to the topic of conversation.

4. **Maxim of Manner** (Be orderly and unambiguous): one should provide information in a way that is orderly and coherent with the context.

Do we always follow these principles? The answer is no. According to Grice, whenever we violate or ***flout*** a maxim, ***implicature*** arises. An implicature is the actual meaning behind an utterance. The implicature must be recovered for the correct purpose to be conveyed. From several psycholinguistic analyses, it has been found that sarcasm arises whenever there is a flouting of the Gricean maxim(s).

Flouting a maxim is different from violating a maxim. Consider a conversation between A and B.

- A: Can you tell me where this road goes?
- B: Can I tell you the name of my new album?

This reply from B violates the maxim of *relevance*.

Now consider another conversation between C and D.

- C: can you tell me where this road goes?
- D: The road goes nowhere, we do!

The reply from D is a flouting of the maxim of relevance and quality. Here D did not violate it as the reply is still relevant and valid. But un-intentionally, it gave rise to the implicature that the information is evident regarding the road which should not be asked (maybe the end of the road was two steps away!)

The main challenge is to implement these theories as mathematical models of representation that can be used in a deep learning setup.

As for our task, we further explored the sarcasm explanation in code-mix ***Hinglish*** conversation. The dataset used is called **WITS**(*Why Is This Sarcastic*) which was curated from the Indian tv show called *"Sarabhai vs Sarabhai"*. The dataset contains both the English and the code-mix explanation for a given sarcastic utterance in a dialogue. The *WITS* dataset was further enriched with the meaning-preserving English translation of all the dialog conversations.

As in the previous work from our research team, the Modality Aware Fusion(**"MAF"**)was incorporated alongside the Multimodal Context Aware Attention(**MCA2**) into the proposed architecture. The MCA2 mechanism helps establish the deep semantic relationship between different modalities of visual and acoustic signals alongside the text modality.

To extend this task further, we are unifying psycho-lingusitic theories along with the deep learning architecture which can help in better inference and understanding of presence of sarcasm and thereby the intended meaning, which can help generate better explanations.

We propose the ***Explanandum*** and the ***Explanans*** as two Major modules. Explanandum, as used in logic and philosophy, means what to explain. In our case, the explanandum part is the one that learns the features required to understand what is the accurate semantic representation of a given dialog. The Explanans part is the one that incorporates the Gricean Maxims with the encoded features from the explanandum. It is responsible for the understanding, observation, and generation of explanations.

In the following sections we present the

1. Motivation
2. Related Work
3. Dataset
4. Proposed Methodology
5. Results
6. Discussion and,
7. Future Scope

Chapter 2

Motivation

One of the most complex problems in NLP is sarcasm recognition. Extracting the implicatures as a subset of sentiment analysis is challenging. It lowers the effectiveness of feedback systems and collaboration in effective communication since the genuine hidden sentiments are bypassed, which is the exact reverse of what the traditional paradigms of sentiment detection were designed to do.

People can now express their thoughts in ways they previously couldn't because to the development of open internet forums, social media platforms, and websites. Their statements and comments reveal their goals, viewpoints, and feelings. Analysis of the emotions underlying their viewpoint is so crucial. It contributes to better affairs and suppresses hostile forces that abuse these platforms to promote false information and disturb social harmony.

Sarcasm does, as has been stated, depend on various things. Context, verbal and nonverbal indicators, the emotional condition of the parties involved, cognitive factors, etc. are all relevant. Even for humans, it might be challenging to decipher and understand those indicators in many situations appropriately. So far, most of the work done on sarcasm detection has been on text data. Multimodal cues like acoustic and visual (picture + video) signals had to be included since they improved context understanding and made it easier to compare the literal and intended meaning.

In human communication, indirect speech, like sarcasm, accomplishes a variety of discourse aims. While the figurative nature of language encourages speakers to attain particular pragmatic goals, it might be difficult for AI agents to understand these quirks of human communication. Although the topic of sarcasm identification has received much attention in dialogue analysis, it is crucial to explain the underlying sarcastic connotation of a discussion in order to fully grasp its underlying meaning and produce suitable answers.

The arrival of large language models has given immense inference mechanisms for most of the NLP tasks. But too much reliance on them leads to a loss of explainability as the fundamental theories of social behavior and language comprehension fall apart.

The relationship between a sound theory of sarcasm comprehension can help us to create better and persona-aware bots that can help understand the sarcasm and then explain it to those who find it difficult to infer the hidden meaning.

Chapter 3

Related Work

Gibbs, in his 1986 paper on the Psycho-linguistics of sarcasm, performed six experiments to gauge the factors leading to comprehension and understanding of irony. Before this work, there was too much emphasis on the classical 'Standard Pragmatic Model', which says that our brain first retrieves the literal meaning of something we perceive. If it encounters any anomaly arising out of the dichotomies and the semantic mismatch, it only processes further and recovers the implicatures, i.e., the true hidden meaning behind the stimuli. The statistical analysis of the participants' response time during sentence reading and comprehension was done. The results showed that any utterance has no fixed meaning. The comprehension and understanding entirely depend on the shared belief between the person(s) involved in the conversation and the underlying context. The short-term memory associated with language comprehension also suggested that the standard pragmatic model rarely comes into the picture to understand sarcasm which established the role of a shared ground or belief and contextual significance of any utterance for its comprehension.

In most of the earlier literature, we do see that Incongruity theory is being used to explain the rationale behind something being funny or not. This, however, has a significant drawback of classifying Surprise, Juxtaposition, Atypical, and Violation as humorous too, as pointed out by Warren et al. The divergence between the expected and unexpected, such as the anti-climax of a story, can also be classified mistakenly as sarcasm. Any atypical scenario or particular violation of our traditional social beliefs can also be labeled as sarcasm. These authors proposed the idea called Benign Violation Theory. Benign violation goes against our beliefs and faith, but at the same time, it is okay, i.e., mild. This is the reason behind the origin of humour. For example, when somebody tickles us, our personal space gets evaded. But, at the same time, it feels normal. This was the primary motivation behind the hypothesis. This idea,

when tested, helped reduce the misclassification rate for humorous vs non-humorous situations.

In a well-organized overview on computational sarcasm, Joshi et al. (2017) elaborated on the pertinent datasets, trends, and problems for automatic sarcasm identification. Early attempts to identify sarcasm relied on single text sources like tweets and reviews (Kreuz and Caucci, 2007; Tsur et al., 2010; Joshi et al., 2015; Peled and Reichart, 2017). These early studies mostly concentrated on using linguistic and lexical characteristics to identify sarcasm's telltale signs (Kreuz and Caucci, 2007; Tsur et al., 2010).

The role of machine learning played a considerable part in classification tasks. But, the early days of such tasks depended primarily on strict paradigms. To determine the presence of sarcasm, sarcasm detection tasks mainly focused on rule-based, hashtag-based, and pragmatics such as punctuation and capitalization of characters. In their paper, Joshi et al.[**?**] describe how the need to incorporate standard linguistic theories is necessary for sarcasm detection. They proposed a Context Incongruity theory for detecting sarcasm. Context Incongruity incorporates both explicit as well as implicit incongruity. For example, "I love being ignored" has both the words 'love' and 'ignored' of opposite polarity, and hence the contradiction is explicit. Whereas, in the sentence 'I love your face so much that I would never stop playing cricket with it', any negative words are absent. But the second phrase considered for the context is negative, so the incongruity is implicit. This paper used tweets as well as comments from discussion forums. The best model gives us an accuracy of about 88.76%. There is an improvement of around 8% in the F1 score. Riloff, Maynard datasets were used, out of which Riloff gave the best accuracy. The major drawback was this study did not consider Inter-sentential incongruity.

Ghosh et al in their research, analyzed that since grammatical errors are commonplace across social media, and most of the classical machine learning models incorporated standard grammatical structures such as Bag of Words and POS tagging, the results were not typically good. They proposed a Neural network semantic model for sarcasm detection comprising CNN-LSTM-DNN layers. The F1 score turns out to be 92%. They had used a Twitter dataset of 41K tweets, with 39K tweets as training data, constituting sarcastic and non-sarcastic labels. The model also performed outstandingly on the standard TSUR and RILOFF datasets with F1 scores of 90.1 % and 88.1% respectively. This study established the semantic power of neural networks. But the proposed model failed to classify similar concepts. It correctly classified "I just love Mondays!" as sarcastic. But I could not do the same for the sentence "Thank god it's Monday!"

Recently, attention-based architectures have been proposed to take use of the links

between and within sentences in texts for effective sarcasm recognition (Tay et al., 2018; Xiong et al., 2019; Srivastava et al., 2020). Tay et al., in their research proposed MIARN(Multi-dimensional Intra-Attention Recurrent Network) to detect intra-sentence incongruity. which earlier neural network architecture failed to do so. The architecture used encodings generated from the text, and parallel processing using LSTM and MIARN was done. The output from those was finally passed to a final decoder layer. This study focused on the way opinions are being mined for sarcasm. Models that were used as baselines were CNN-LSTM-DNN and GRNN. Contrastive sentiments were also detected in the proposed architecture. They achieved the best accuracy using MIARN, which came out to be 86.47%.

As it has been said, sarcasm does depend on many factors. It depends on the context, the verbal and the non-verbal cues, the emotional state of involved entities, the cognitive factors, etc. And often, it's not easy to interpret and comprehend those signs correctly, even by human beings. Most of the sarcasm detection work to date was done on text data. Incorporating multimodal cues such as auditory and visual (image + video) cues was necessary as they better help understand the context; thus, the contrast between the literal and intended meaning can easily be extrapolated. Castro et al.[**?**] proposed a new dataset called Multimodal Sarcasm Detection Dataset (MUStARD). It is compiled from multiple popular tv shows with audio-visual utterances and the sarcasm label.

Multimodality helps provide additional cues such as prosody, stress on discourse markers, facial expressions, etc. The models trained over this dataset had shown a performance improvement and reduced the error in the F1 score by 12.9%. This work provided a new challenge of finding the incongruity among modalities, detecting sarcasm in a conversational context, and prominent speaker localization.

Earlier models used concatenation of features extracted from different modalities. Cai et al. collected Twitter data (which consists of images and texts) for their analysis. This study considered the image, attributes, and text as three separate modalities. These modalities were used to create a single feature vector. The final vector was used for classification. It performed better than the baseline with individual modalities and the concatenated features. The researcher used Bidirectional LSTM to extract text from the image. A new dataset had been curated for multimodal bidirectional LSTM. There was no public dataset; hence, the researcher built their dataset in this research paper. CNN, such as ResNet, was used to extract image-specific features. The best overall accuracy came out to be 83.44%.

Bedi et al. curated a dataset called MaSaC from the popular tv show 'Sarabhai vs Sarabhai'. They addressed two areas of code-mixed conversation and humour detection alongside sarcasm detection. The rationale was that most of our conversations switched

back and forth among different languages for the former. For example, some of us use "Hinglish" in our daily conversations. The challenges to extracting meaning from such code mixed dialogues were explained in the paper. The latter hypothesis was based on the reasoning that both the parts are related in the problem space and depend on the context of the situation to extract the hidden semantics. They should be done as a joint task. The modalities used for the processing were textual and acoustic. Visual modality was not considered, introducing noise and hampering the learning process. The proposed model MSH-COMICS for classifying utterances as sarcastic and non-sarcastic and humorous or non-humorous showed promising results.

In the context of conversational AI, analysis of figurative language has also received substantial study. Attention-based RNNs were used by Ghosh et al. (2017) to recognise sarcasm in the presence of context. For the two inputs (sentence and context), two different LSTMs-with-attention were trained, and during prediction, their hidden representations were mixed. Beyond the English language, sarcasm identification research has also been conducted. Sarcasm was identified using rule-based techniques by Bharti et al.(2017) using a corpus of 2000 sarcastic tweets in Hindi. Swami et al. (2018) employed n-gram feature vectors with different ML models to detect sarcasm in a dataset of 5000 humorous Hindi-English code-mixed tweets. Other noteworthy research focus on Italian (Cignarella et al., 2018), Spanish (Ortega-Bueno et al., 2019), Arabic (Abu Farha and Magdy, 2020), and Spanish (Ortega-Bueno et al., 2019).

While sarcasm identification has been the main focus of computational sarcasm studies, some attempts into other areas of figurative language analysis have been explored. Deep learning has been used by Dubey et al. (2019) to translate sarcastic statements into non-sarcastic interpretations. Another method for generating sarcasm was developed by Mishra et al. (2019) by producing context incongruity by fact removal and incongruous phrase substitution. After that, Chakrabarty et al. (2020) introduced an unsupervised framework for snark creation that is built on retrieve-and-edit. Their suggested model uses semantic incongruity and valence reversal to produce sarcastic phrases from their non-sarcastic counterparts.

Also the analysis of statistical experiments has been done on Gricean maxims and their violation of humour and irony, the experiments have never been concrete, and there are not any concrete mathematical models to represent those maxims astutely. Also, the scarce amount of instances curated during these psychological experiments are not apt for deep learning as they are highly prone to overfitting.

A lot of research has been done on detecting sarcasm, but very little, if any, has been done on elucidating the irony of sarcasm. In order to close this gap, this study offers a novel problem formulation by incorporating the psycho-linguistic facets of sarcasm comprehension.

Chapter 4

Dataset

"Sitcoms" or situational comedy, accurately portray human behaviour and mannerism in typical, daily situations. As a result, the NLP research community has utilised such data for sarcasm recognition with success.

The dataset used for the task is **WITS**(Why Is This Sarcastic). WITS is an extension of **MaSaC**, a sarcasm detection dataset with the addition of code-mix and english explanation of the sarcastic utterance in a dialogue. The dataset is a compilation of sarcastic dialogues from a popular Indian TV show called *"Sarabhai vs Sarabhai*. Along with the textual transcripts of the conversations, the dataset also contains multimodal signals of audio and video.

The TV show's original dataset had 45 worth of episodes, but WITS has 10 additional episodes as well with their transcriptions and audio-visual limits. Then, the sarcastic utterances from this updated dataset were chosen, and the utterances that should be included in the dialogue context for each of them were manually defined. Finally, the dataset is left with 2240 sarcastic dialogues with the number of contextual utterances ranging from 2 to 27. Each of these instances is manually annotated with a corresponding natural language explanation interpreting its sarcasm.[26] Each explanation contains four primary attributes – source and target of sarcasm, action word for sarcasm, and an optional description for the satire as illustrated in Figure **??**.

Each instance of the text data conists of

1. Episode number
2. Episode name
3. Context speakers
4. Target utterance

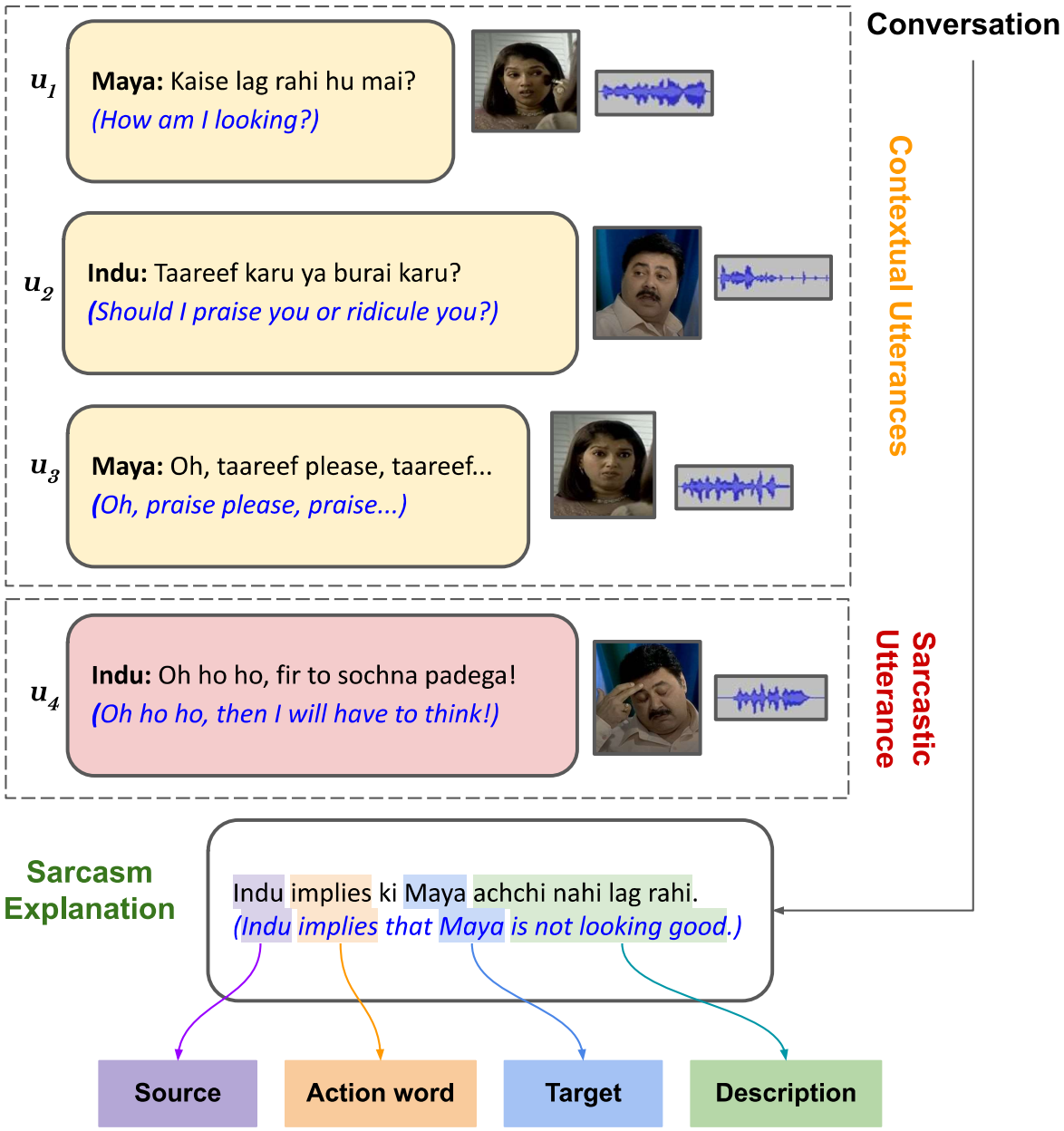


Figure 4.1: Sarcasm Explanation in Dialogues (). Given a sarcastic dialogue, the aim is to generate a natural language explanation for the sarcasm in it. *Blue text represents the English translation for the text.*

5. Context utterances
6. Target Speaker
7. Sarcasm target
8. Sarcasm type (Intended or Perceived)
9. Code-Mix Explanation
10. English Explanation

The dataset as desrcribed in 4.2 was further extended in this work by adding semantic preserving English translation of all the context utterances and the target utterance.

For the respective visual and acoustic modalities, features were extracted from the video and audio file tools described in the Results section.

We further developed 3 attributes for our dataset. For each of the instances, we developed

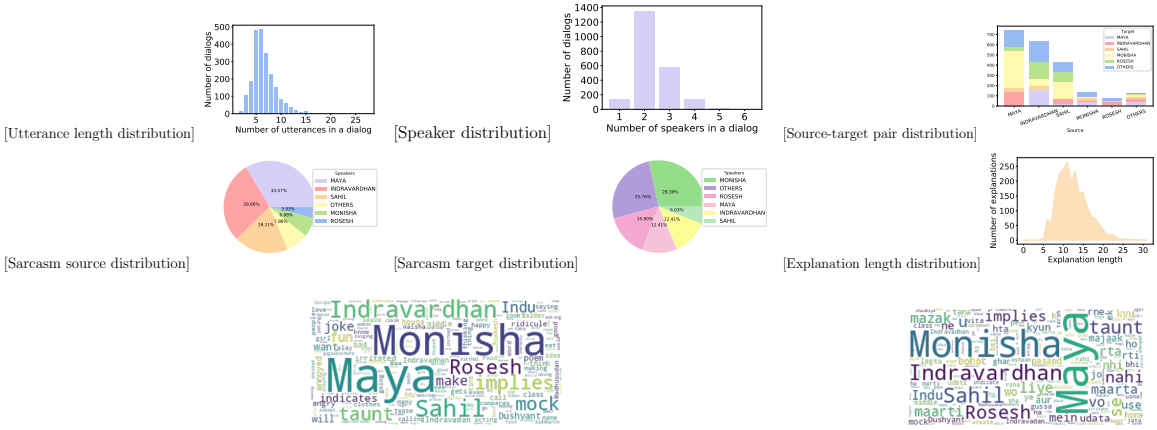
1. **Simple Graph**: Each sentence in the dialog is represented as node. It is a complete unweighted graph.
2. **Entailment Graph**: Each sentence in the graph is connected to another sentence with associated entailment value. For every pair of nodes x and y, we have two edges from x to y and y to x.
3. **Contradiction Graph**: Each sentence in the graph is connected to another sentence with an associated contradiction value. For every pair of nodes x and y, we have two edges from x to y and y to x.

The respective entailment and contradiction values were computed using the Facebook **MNLI** pre-trained model in a zero-shot setup.

The data was split into training, validation, and test sets in the ratio of 80:20:10. the training set thus, have 1792 instances and the remaining two have 224 instances each.

Table 4.1: Satistics of the dataset

# Dialogues	**# Utterances**	**# English utts**	**# Hindi utts**	**# Code-mixed utts**	**Avg. utt/dlg**
2240	9080	101	1453	7526	4.05
Avg. speaker/dlg	**Avg. words/utt**	**Avg. words/dlg**	**Vocab size**	**English vocab size**	**Hindi vocab size**
2.35	14.39	58.33	10380	2477	7903



[Utterance length distribution] [Speaker distribution] [Source-target pair distribution]

[Sarcasm source distribution] [Sarcasm target distribution] [Explanation length distribution]

[Word cloud for English words] [Word cloud for code-mixed words]

Figure 4.2: Distribution of attributes in the dataset. The number of utterances in a dialog lies between 2 and 27. Maximum number of speakers in a dialogue are 6. The speaker 'Maya' is the most common common sarcasm source while the speaker 'Monisha' is the most prominent sarcasm target.

Chapter 5

Proposed Methodology

5.1 Foundation architecture

The current work is based upon the architecture developed by Kumar et al.[26] Their primary goal was to smoothly integrate multimodal knowledge into the BART architecture. To this end, they introduced *Multimodal Aware Fusion*, an adapter-based module that comprises of *Multimodal Context-Aware Attention (MCA2)*() and *Global Information Fusion (GIF)* [26]mechanisms. Given the textual input sarcastic dialogue along with the audio-video cues, the former aptly introduces multimodal information in the textual representations, while the latter conglomerates the audio-visual information infused textual representations. This adapter module can be readily incorporated at multiple layers of BART/mBART to facilitate various levels of multimodal interaction. Figure 5.1 illustrates that model architecture.

5.1.1 Multimodal Context Aware Attention

Textual representations interact directly with other modalities when using the conventional dot-product-based cross-modal attention strategy. Here, the multimodal representations serve as the key and value while the text representations serve as the query against them. A direct merging of multimodal information might not retain all contextual information because each modality originates from a distinct embedding subspace and might also leak a lot of noise into the final representations.

Prior to performing the conventional scaled dot-product attention, the generation of multimodal information-conditioned key and value vectors is carried out. The description given by Kumar et al[26]. of the procedure is as follows:

Given the intermediate representation H generated by the GPLMs at a specific

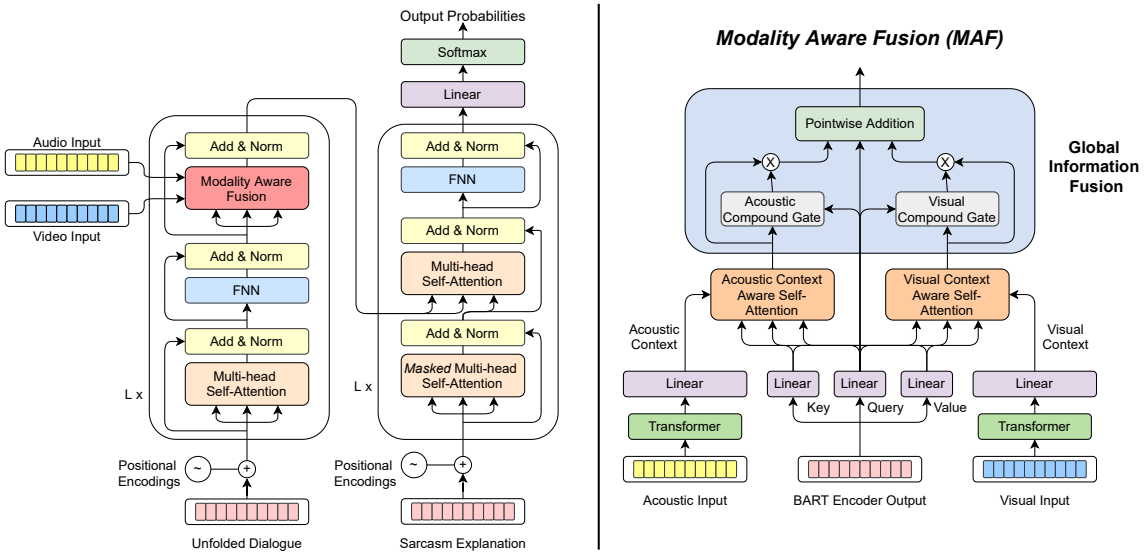


Figure 5.1: Model architecture for . The Foundational proposed Multimodal Fusion Block captures audio-visual cues using Multimodal Context Aware Attention (MCA2) which are further fused with textual representations using Global Information Fusion (GIF) block.

layer, the query, key, value and value vectors Q, K, and $V \in R^{n \times d}$, respectively, as given in Equation 5.1, where W_Q, W_K, and $W_V \in R^{d \times d}$ are learnable parameters. Here, n denotes the maximum sequence length of the text, and d denotes the dimensionality of the GPLM generated vector.

$$\left[QKV\right] = H\left[W_Q W_K W_V\right] \tag{5.1}$$

$$\begin{bmatrix}\hat{K} \\ \hat{V}\end{bmatrix} = (1 - \begin{bmatrix}\lambda_k \\ \lambda_v\end{bmatrix}) \begin{bmatrix}K \\ V\end{bmatrix} + \begin{bmatrix}\lambda_k \\ \lambda_v\end{bmatrix} (C \begin{bmatrix}U_k \\ U_v\end{bmatrix}) \tag{5.2}$$

Finally, the multimodal information infused vectors are used to compute the traditional scaled dot-product attention.

5.1.2 Global Information Fusion

In order to combine the information from both the acoustic and visual modalities, Kumar et al. [26] design the GIF block. They proposed two gates, namely the *acoustic*

gate (g_a) and the *visual gate* (g_v) to control the amount of information transmitted by each modality.

5.2 Gricean Maxims

Grice asserted that people in a conversation follow the four ***maxims*** of conversation for efficient communication. These maxims are

1. ***Maxim of Quality***: Don't share anything that you know to be untrue or illogical; instead, make your contribution true.
2. ***Maxim of Quantity***: Be as informative as required. No less, no more.
3. ***Maxim of Relevance***: Be relevant
4. ***Maxim of Manner***: Be observant, thus steer clear of obscurity and ambiguity and aim for brevity and organisation.

These maxims could be understood by a simple example. Suppose Raju needs to build a boat. His brother will hand him a hammer instead of a sword(being relevant), give him more than a single block of wood(maxim of quanity), good wood rather than rotten one(maxim of quantity) and will try to provide it as quickly and efficiently as possible(maxim of manner). These were seen by Grice as examples of general laws regulating reasonable, cooperative behaviour rather than as arbitrary conventions.

Whenever a particular maxim is flouted, it gives rise to *implicatures*. Implicature is the intended meaning beyond the said statement. The theoretical account of conversational implicature for a speaker S and the listener L can be given as:

S conversationally implicates m *iff* S implicates m when:

- S is presumed to be observing the Cooperative Principle. It is called ***Cooperative Presumption***
- The supposition that S believes m is required to make S's utterance consistent with the Cooperative Principle. It is called ***Determinacy***
- S believes (or knows), and expects L to believe that S believes that L is able to determine the consistency is true (Mutual Knowledge and Theory of Mind).

At the heart of Gricean theory lies the ***Calculatibility principle.*** It states that it must be possible to resolve conversational implicatures. The hearer will use the following information to determine whether a particular conversational implicature is present:

the conventional meaning of the words used, together with the identity of any references that may be involved;

the Cooperative Principle and its maxims;

the context, linguistic or otherwise, of the utterance;

Background knowledge

the fact (or alleged reality) that both participants have access to all relevant things falling under the aforementioned headings and both participants are aware of or believe this to be the case.

Given the dependence of implicature on the the three conditions of *Presumption*, *Determinacy* and, *Mutual shared knowledge*, He also postulated ***Grice's Razor***, similar to the ***Occam's Razor***. It states that *"senses are not be multiplied beyond necessity."* It is theoretically more cost-effective to explain and anticipate an event in terms of psycho-social principles rather than positing senses and the like, which cannot be thus described.

Sarcasm in conversational dialog involves implicatures and oftentimes the maxims of cooperative principles are flouted to express sarcasm.

5.2.1 Explanans

Explanans is a latin term which means *'what to explain'*. Our work incorporates the node2vec embedding from the Simple Graphs as a new modality. This simple graph is analogy to the chaotic relationship between different sentences that need to be resolved to get atleast a proper semantic representation which may or may not be completely true. The encoder from the foundation architecture gives the representation that has to be fed into the *Explanandum* .

5.2.2 Explanandum

Explanandum means *'What does the explaining'*. In our task, The explanandum comprises of:

1. Commonsense knowledge graph: To get commonsense knowledge regarding the current dialog scenario. Node2Vec embeddings are used as a feature for further processing

2. Maxim-Quality: It is the feature extracted from Node2Vec embeddings of the Contradiction graph. The embeddings are passed through a transformer encoder.

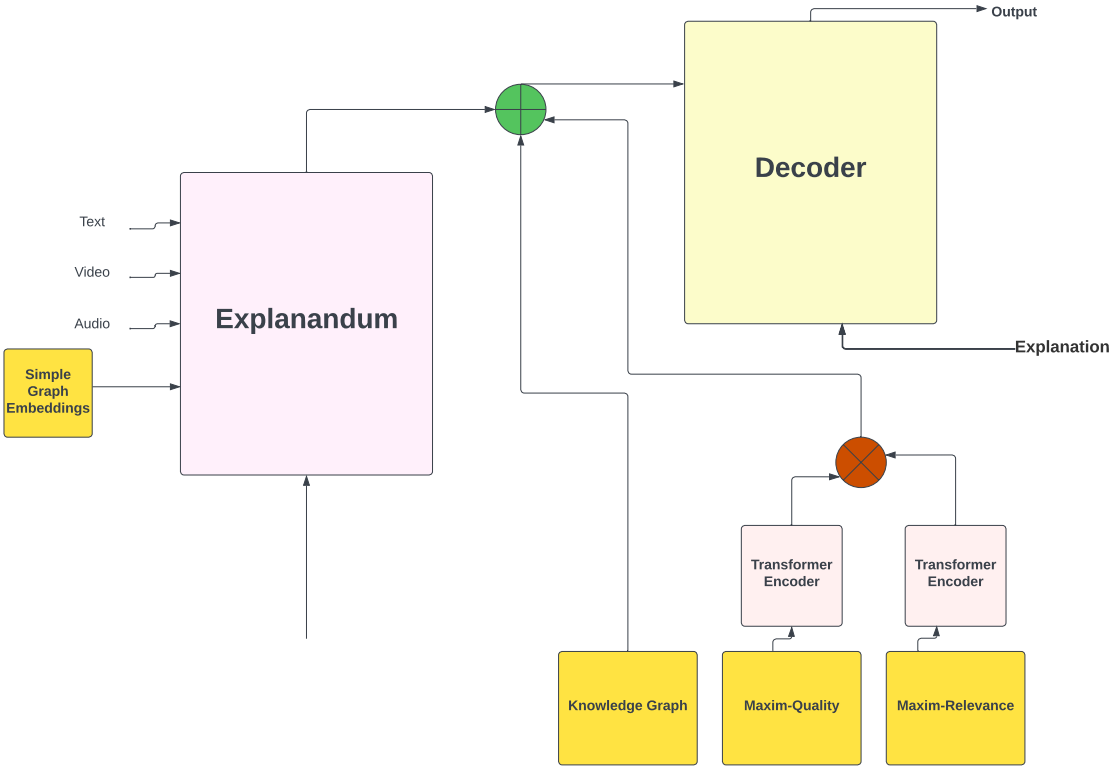


Figure 5.2: The Proposed architecture. Explanans comprises of modules outside the Explanandum.

3. Maxim-Relevance: It is the feature extracted from Node2Vec embeddings of the Entailment graph. The embeddings are passed through a transformer encoder.

4. Decoder.

The encoded representation from the previous steps had to be reconciled to feed to the final decoder. We applied an attention layer over the representation from Maxim-Quality and the Maxim-Relevance. To all the final representations, we performed adding the embeddings to get the final embedding representation which was fed to the decoder.

5.3 Results

We first implemented the foundation architecture for our task which is referred to in the 5.1. For the acoustic and visual modality, we had the extracted features with us.

Networkx module was used to create unweighted graph for Simple Graph representation. For the Entailment and Contradiction graph, we used Facebook's **MNLI** to generate the degree of entailment and contradiction between a pair of sentences.

These graphs were further passed to Node2Vec for embedding generation.

5.3.1 Baselines

As textual data has been the prime modality, several baseline were implemented:

1. **RNN**: The OpenNMT version of the seq-2-seq architecture was used.
2. **Transformer**: The standard encoder-decoder architecture was implemented
3. **Pointer Generator Network**: It is used for seq-2-seq task as well as for summarization.
4. **BART**: It is a denoising autoencoder model with standard machine translation architecture with a bidirectional encoder and an auto-regressive left-to-right decoder. its base version was used.

   The results f english text only dataset is found in 5.1.

		R1	R2	RL	B1	B2	B3	B4	METEOR
5*Code-mixed	RNN	8.94	0.98	8.61	1.24	0.29	1.63	1.22	3.30
	Transformer	3.20	0.25	3.19	0.40	0.01	1.15	8.20	1.11
	PGN	23.37	4.83	17.46	17.32	6.68	1.58	0.52	23.54
	BART	**41.49**	16.66	**38.81**	**33.27**	18.00	10.87	6.71	31.88
	mBART	40.82	17.35	38.08	32.75	18.33	**11.8**	**8.31**	31.12
5*English	RNN	8.94	0.98	8.61	1.24	0.29	1.63	1.22	3.30
	Transformer	3.20	0.25	3.19	0.40	0.01	1.15	8.20	1.11
	PGN	23.37	4.83	17.46	17.32	6.68	1.58	0.52	23.54
	BART	**41.49**	16.66	**38.81**	**33.27**	18.00	10.87	6.71	31.88
	mBART	40.82	17.35	38.08	32.75	18.33	**11.8**	**8.31**	31.12
	MAF-TAVB	36.69	17.10	37.37	33.20	18.69	**12.37**	**8.58**	30.40
	Proposed	39.01	17.20	36.7	31.00	17.54	**11.3**	**6.9**	30.07

Table 5.1: Multilinguality

5.3.2 Evaluation Metrics

- ROUGE
  Recall-Oriented Understudy for Gisting Evaluation, or ROUGE, is an evaluation metric. In reality, ROUGE is a collection of measures rather than a single one. The metrics contrast an automatically generated summary or translation with a reference summary or translation, or a collection of references. ROUGE is solely based on memory.
  The following are the five evaluation criteria for ROUGE:

  1. ROUGE-N: N-gram overlap between the reference and system summaries. There are two subgroups of ROUGE-N: ROUGE-1 and ROUGE-2. The overlap of unigrams between the system and reference summaries is referred to as ROUGE-1. The bigram overlap between the system and reference summaries is known as ROUGE-2.
  2. ROUGE-L: Using statistics based on the longest common sequence. The longest common subsequence issue automatically recognizes the longest co-occurring in sequence n-grams while naturally accounting for sentence-level structure similarities.
  3. ROUGE-W: LCS-based statistics that are weighted in favor of consecutive LCSes.
  4. ROUGE-S: Statistics for co-occurrences based on skip-bigram. Any pair of words in their sentence order is a skip-bigram.
  5. ROUGE-SU: Unigram-based co-occurrence statistics combined with skip-bigram.

- BLEU The Bi-Lingual Evaluation Understudy (BLEU) compares the machine-generated translation, also referred to as the candidate translation, with existing human-generated translations, also referred to as the reference translations. The value of the BLEU score, like any precision-based statistic, is always a number between 0 (worst) and 1. (best). Modified n-gram precision and best match length in BLEU are used to approximate precision and recall, respectively.

- METEOR It is utilized to assess how closely the candidate text produced by an ML model corresponds to the reference text that is supposed to be generated. When assessing a match, METEOR considers both precision and recall. METEOR alters the accuracy and recall computations with a weighted F-score based on mapping unigrams and a penalty function for erroneous word order.

An example Explanation Generation for different Models can be seen in the figure below.

SAHIL: Ab tumne ghar ki itni saaf safai ki hai and secondly us Karan Verma ke liye pasta, lasagne, caramel custard banaya. *Now you have cleaned the house so much and secondly made pasta, lasagne, caramel custard for that Karan Verma.*
MONISHA: Walnut brownie bhi. *And walnut brownie too.*
SAHIL: Walnut brownie, matlab wo khane wali? *You mean edible walnut brownie?*

Gold	Sahil monisha ki cooking ka mazak udata hai *Sahil makes fun of Monisha's cooking.*
BART	Monisha sahil ko walnut brownie ki matlab wo khane wali. *Walnut Brownie to Monisha Sahil means she eats*
MAF-TAV$_B$	Sahil monisha ki cooking ka mazak udata hai *Sahil makes fun of Monisha's cooking.*

Chapter 6

Discussion

As per the results in the above-given Table-2, the MAF-TAV performance still had been superior. Multimodality definitely helps improve the task of explanation generation. Our proposed method of Explanans and Explanandum still did good. To compare whether it actually is understanding the sarcasm, We tried to predict the class of utterance as Intended, Perceived or both. As per the results, the encoded representation from the explanans improved this classification task by 7%. This suggests that, our proposed architecture of explanandum is inferring the implicature and the semantics more accurately. This inference is what we need for a step towards artifical general intelligence

Chapter 7

Future Scope

7.1 Psychology and Sarcasm

As evident from the results, psychological theories of language comprehension do play a crucial role. Although Gricean maxims does play a crucial role to understand the hidden meaning, it does fail in many scenarios such as the case of deliberately *opting out* of following a maxim.

There have been new theories such as the *relevance theory* and the *neo-gricean theory*. These may help overcome the various challenges the cooperative principle faces.

7.2 Cognition and sarcasm

Mishra et al. proposed that the cognitive load while processing the semantic incongruity during sarcasm comprehension must be higher. Since gaze movement is correlated with such cognitive processing, it must be of some use in sarcasm detection. The eye-movement patterns of the human reader were used to extract the mental feature to enhance the overall feature space. In this paper, the authors took note of the incongruity theory and made specific observations and hypotheses, further supported by their results. Sarcasm requires more processing time. Hence it is reasonable to deduce the longer fixation duration would indicate sarcasm. They framed a dataset created using the Gaze Pattern of human readers. This was further used to generate line graphs called Scanpaths. Every node denoted the fixation duration, and the edge characterized the saccadic movement between the words.

Bibliography

[1] Ibrahim Abu Farha and Walid Magdy. From Arabic sentiment analysis to sarcasm detection: The ArSarcasm dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39, Marseille, France, May 2020. European Language Resource Association.

[2] Salvatore Attardo, Jodi Eisterhold, Jennifery Hay, and Isabella Poggi. Multimodal markers of irony and sarcasm. *Humor: International Journal of Humor Research*, 16(2), 2003.

[3] Manjot Bedi, Shivani Kumar, Md Shad Akhtar, and Tanmoy Chakraborty. Multimodal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing*, pages 1–1, 2021.

[4] Santosh Kumar Bharti, Korra Sathya Babu, and Sanjay Kumar Jena. Harnessing online news for sarcasm detection in hindi tweets. In B. Uma Shankar, Kuntal Ghosh, Deba Prasad Mandal, Shubhra Sankar Ray, David Zhang, and Sankar K. Pal, editors, *Pattern Recognition and Machine Intelligence*, pages 679–686, Cham, 2017. Springer International Publishing.

[5] Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy, July 2019. Association for Computational Linguistics.

[6] 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.

[7] Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. Rˆ3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7976–7986, Online, July 2020. Association for Computational Linguistics.

[8] Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4351–4360, Online, July 2020. Association for Computational Linguistics.

[9] Dushyant Singh Chauhan, Gopendra Vikram Singh, Navonil Majumder, Amir Zadeh, Asif Ekbal, Pushpak Bhattacharyya, Louis-philippe Morency, and Soujanya Poria. M2h2: A multimodal multiparty hindi dataset for humor recognition in conversations. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, ICMI '21, page 773–777, New York, NY, USA, 2021. Association for Computing Machinery.

[10] Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, Paolo Rosso, et al. Overview of the evalita 2018 task on irony detection in italian tweets (ironita). In *Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*, volume 2263, pages 1–6. CEUR-WS, 2018.

[11] Herbert L. Colston. Salting a wound or sugaring a pill: The pragmatic functions of ironic criticism. *Discourse Processes*, 23(1):25–45, 1997.

[12] Herbert L Colston and Shauna B Keller. You'll never believe this: Irony and hyperbole in expressing surprise. *Journal of psycholinguistic research*, 27(4):499–513, 1998.

[13] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[14] Abhijeet Dubey, Aditya Joshi, and Pushpak Bhattacharyya. Deep models for converting sarcastic utterances into their non sarcastic interpretation. In *Proceedings*

of the ACM India Joint International Conference on Data Science and Management of Data, CoDS-COMAD '19, page 289–292, New York, NY, USA, 2019. Association for Computing Machinery.

[15] Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016.

[16] Debanjan Ghosh, Alexander R. Fabbri, and Smaranda Muresan. Sarcasm analysis using conversation context. *Computational Linguistics*, 44(4):755–792, December 2018.

[17] Debanjan Ghosh, Alexander Richard Fabbri, and Smaranda Muresan. The role of conversation context for sarcasm detection in online interactions. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 186–196, Saarbrücken, Germany, August 2017. Association for Computational Linguistics.

[18] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[19] Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. Humor knowledge enriched transformer for understanding multimodal humor. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12972–12980, May 2021.

[20] Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China, November 2019. Association for Computational Linguistics.

[21] Stacey L. Ivanko and Penny M. Pexman. Context incongruity and irony processing. *Discourse Processes*, 35(3):241–279, 2003.

[22] Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. Automatic sarcasm detection: A survey. *ACM Comput. Surv.*, 50(5), September 2017.

[23] 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.

[24] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017.

[25] Roger Kreuz and Gina Caucci. Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 1–4, Rochester, New York, April 2007. Association for Computational Linguistics.

[26] 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, 2022.

[27] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*, 2018.

[28] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018.

[29] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[30] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[31] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

[32] Abhijit Mishra, Tarun Tater, and Karthik Sankaranarayanan. A modular architecture for unsupervised sarcasm generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6144–6154, Hong Kong, China, November 2019. Association for Computational Linguistics.

[33] Henri Olkoniemi, Henri Ranta, and Johanna K Kaakinen. Individual differences in the processing of written sarcasm and metaphor: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(3):433, 2016.

[34] Shereen Oraby, Vrindavan Harrison, Amita Misra, Ellen Riloff, and Marilyn Walker. Are you serious?: Rhetorical questions and sarcasm in social media dialog. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 310–319, Saarbrücken, Germany, August 2017. Association for Computational Linguistics.

[35] Reynier Ortega-Bueno, Francisco Rangel, D Hernández Farıas, Paolo Rosso, Manuel Montes-y Gómez, and José E Medina Pagola. Overview of the task on irony detection in spanish variants. In *Proceedings of the Iberian languages evaluation forum (IberLEF 2019), co-located with 34th conference of the Spanish Society for natural language processing (SEPLN 2019). CEUR-WS. org*, volume 2421, pages 229–256, 2019.

[36] Shruti Palaskar, Jindrich Libovický, Spandana Gella, and Florian Metze. Multimodal abstractive summarization for how2 videos, 2019.

[37] Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1383–1392, Online, November 2020. Association for Computational Linguistics.

[38] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[39] Lotem Peled and Roi Reichart. Sarcasm SIGN: Interpreting sarcasm with sentiment based monolingual machine translation. In *Proceedings of the 55th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1690–1700, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[40] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy, July 2019. Association for Computational Linguistics.

[41] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[42] Richard M. Roberts and Roger J. Kreuz. Why do people use figurative language? *Psychological Science*, 5(3):159–163, 1994.

[43] Patricia Rockwell. Vocal features of conversational sarcasm: A comparison of methods. *Journal of psycholinguistic research*, 36(5):361–369, 2007.

[44] 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, 2020.

[45] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[46] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*, 2020.

[47] Guangyao Shen, Xin Wang, Xuguang Duan, Hongzhi Li, and Wenwu Zhu. Memor: A dataset for multimodal emotion reasoning in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 493–502, New York, NY, USA, 2020. Association for Computing Machinery.

[48] Himani Srivastava, Vaibhav Varshney, Surabhi Kumari, and Saurabh Srivastava. A novel hierarchical BERT architecture for sarcasm detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 93–97, Online, July 2020. Association for Computational Linguistics.

[49] Sahil Swami, Ankush Khandelwal, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. A corpus of english-hindi code-mixed tweets for sarcasm detection. *arXiv preprint arXiv:1805.11869*, 2018.

[50] Sabina Tabacaru and Maarten Lemmens. Raised eyebrows as gestural triggers in humour: The case of sarcasm and hyper-understanding. *The European Journal of Humour Research*, 2(2):11–31, Oct. 2014.

[51] Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su. Reasoning with sarcasm by reading in-between. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1010–1020, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[52] Oren Tsur, Dmitry Davidov, and Ari Rappoport. Icwsm — a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. *Proceedings of the International AAAI Conference on Web and Social Media*, 4(1):162–169, May 2010.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[54] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[55] Henry M Wellman. *Making minds: How theory of mind develops*. Oxford University Press, 2014.

[56] Tao Xiong, Peiran Zhang, Hongbo Zhu, and Yihui Yang. Sarcasm detection with self-matching networks and low-rank bilinear pooling. In *The World Wide Web Conference*, WWW '19, page 2115–2124, New York, NY, USA, 2019. Association for Computing Machinery.

[57] Nan Xu, Zhixiong Zeng, and Wenji Mao. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3777–3786, Online, July 2020. Association for Computational Linguistics.

[58] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer, 2021.

[59] Baosong Yang, Jian Li, Derek F. Wong, Lidia S. Chao, Xing Wang, and Zhaopeng Tu. Context-aware self-attention networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):387–394, Jul. 2019.

[60] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2019.

[61] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.