Affect-aware Conversational System

Asif Ekbal
AI-NLP-ML Research Group
Department of Computer Science and Engineering
IIT Patna, Patna, India
Email: asif.ekbal@gmail.com, asif@iitp.ac.in

CEP on Deep Learning for NLP
IIT Patna

Jan 21, 2020

Acknowledgement: Help from Maujama

Outline

- Background: Conversational AI
- Background: NLG

- Inducing Courteousness in Customer Care Responses

- Multi-modal NLG
  - Attribute-aware and Position-aware Deep Neural Framework for NLG
  - Sentiment and Emotion Controlled NLG

- Summary and Conclusion

Artificial Intelligence and Conversational Agents

- ***Artificial intelligence (AI)*** is one of most-discussed technology topics among the researchers, consumers and enterprises today

- **Conversational AI powered by NLP and ML** has been in the centre of AI revolution during the last few years

Examples: Conversational AI Systems

Phone-based Personal Assistants

SIRI, Cortana, Google Now

Talking to your car

Communicating with robots

Clinical uses for mental health

Chatting for fun

The most simplest form of Conversational System: ***Chatbot***

Intelligent Assistants with current percentage of users: Major Players

- **Microsoft Cortana**: 49%
- **Apple Siri**: 47%
- **Google Assistant**: 23%
- **Amazon Alexa** : 13%

- As part of Statista study
  - Over 64% of business respondents believe that chatbots allow them to provide a more personalized service experience for customers
  - In the field of e-commerce, this is more than the 34% (as reported in 2017)

Chatbot: Impact

- **Chatbot Report 2019: Global Trends and Analysis** at the **Chatbots Magazine** platform (https://chatbotsmagazine.com/chatbot-report-2019-global-trends-and-analysis-a487afec05b)

  - Business Insider experts predict that by 2020, 80% of enterprises will use chatbots
  - By 2022, banks can automate up to 90% of their customer interaction using chatbots

  - According to Opus Research, by 2021, 4.5 billion dollars will be invested in chatbots

More and more organizations are moving toward a conversation-driven interface to better engage their customers and drive efficiency through automation.

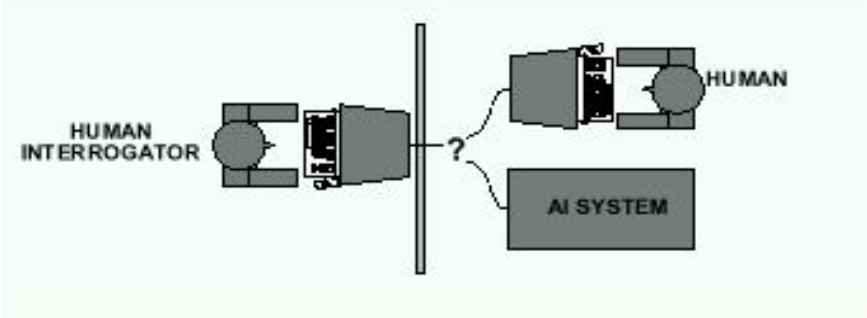
Where is the beginning of Conversational AI?

Alan Turing Test::
Conversational Model (Question-Answering types)

Alan Turing's 1950 article ***Computing Machinery and Intelligence***

Acting Humanly: The Full Turing Test

- "Can machines think?" ←→ "Can machines behave intelligently?"
- The Turing test (The Imitation Game): Operational definition of intelligence



- Computer needs to posses: Natural language processing, Knowledge representation, Automated reasoning, and Machine learning
- Problem: 1) Turing test is not reproducible, constructive, and amenable to mathematic analysis. 2) What about physical interaction with interrogator and environment?
- Total Turing Test: Requires physical interaction and needs perception and actuation.

*

History: Chatbot

- ELIZA
  - Developed in the 1960s at MIT Lab
  - *A way to demonstrate the superficiality of human-to-machine communication by matching user prompts with simple scripted responses*
  - Looks for pronouns and verbs
  - 'You' becomes 'I' and vice versa

An example:

User: You are a dork.

ELIZA: What makes you think I am a dork?

History: Chatbot

- **PARRY (1972)**
  - Was put through the wringer in the early 1970's by a group of 33 psychologists using a variation of the "Turing Test" (of Imitation Game fame)
  - It succeeded in fooling its human examiners 52% of the time

- **RACTER (1984)**
  - Designed in tongue-cheek manner
  - Uses remarkably minimal resources
  - Plays a very active, almost aggressive role, jumping from topic to topic
  - Entertain its users until boredom occurs

History: Chatbot

- ALICE (1995): https://www.chatbots.org/chatbot/a.l.i.c.e/
  - Artificial Linguistic Internet Computer Entity
  - A free software chatbot created in AIML (Artificial Intelligence Markup Language), an open, minimalist, stimulus-response language for creating bot personalities
  - Three time loebner prize winner
  - Developed by Richard Wallace

Today's Chatbot: A Long way from ELIZA

- Nowadays, ***Chatbots*** have grown into a full-blown industry with constant innovations bridging the human-to-machine communication gap
  - *Going beyond simple tasks like playing a song or booking an appointment*

- Beyond knowledge-based conversational agents that match a query to a predefined set of answers

- Chatbot should mimic the dynamics of human conversations

BUT how?

Today's Chatbot: A Long way from ELIZA

- **Generating coherent and engaging responses in conversations**
  - Through Deep Language Understanding and Reasoning
- *Should understand a user's need, context and mood*
- *Should be able to respond with personalization, sentimental and emotional analysis*
- **Advanced NLP and ML Systems**
  - Beyond understanding a single sentence or taking discrete actions
  - Understanding long-form sentences in specific contexts
  - Balancing human-like aspects such as ***specificity*** and ***empathy***

Empowering AI for Human-like Conversation

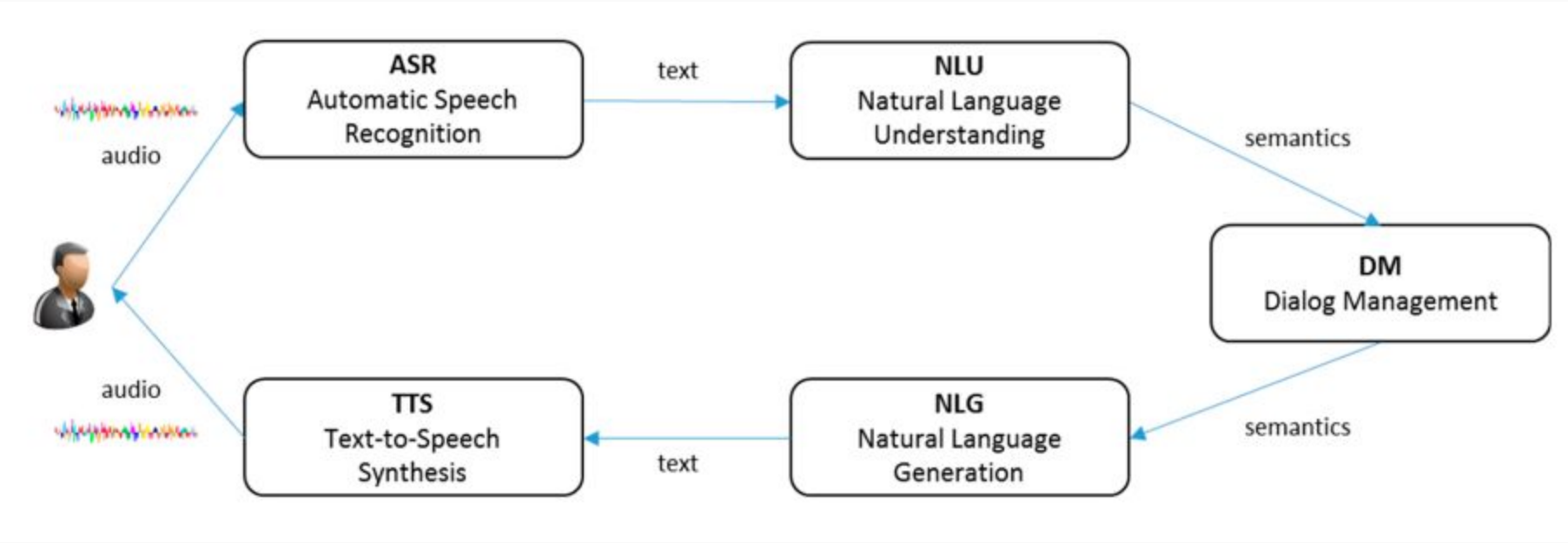
- **AI has to master the art of conversation at human level, then it has an uphill task ahead (*Facebook AI*)**
  - Consistency
  - Specificity
  - **Empathy:**

    Affect-awareness (Sentiment-aware, Emotion-aware), Courteousness etc.
  - Knowledgeability, and
  - Multimodal understanding

Dialogue Agents: Types

- **Open Chit-Chat Agents (Open IE)**
  - Designed for extended conversations, set up to mimic the unstructured conversational or ‘chats’ characteristic of human-human interaction
  - NOT focused on a particular task like airline reservation etc.
  - Systems often have an entertainment value, such as *Microsoft’s XiaoIce*
- **Task-oriented Dialog Agents**
  - Designed for a particular task and set up to have short conversations to get information from the user to help complete the task
  - E.g. Digital assistants like Siri, Cortana, Alexa, Google Now/Home, etc.
  - Agents can give travel directions, control home appliances, find restaurants, or help make phone calls or send texts

Modules in a Task-Oriented Conversational Agent



Natural Language Generation

- One of the key components to a dialogue system
- Goal of NLG is to generate natural language sentences given the semantics
  - Often performed in two steps, *viz.* *Content Planning* and *Sentence Realization*
- Content Planning: by the *Dialogue manager* (“*what to say*”)
- Sentence Realization: *how to say it*

For e.g.,

User: **Book a flight from Kolkata to Delhi.**

System: **Can you please specify the date of travel?**

Existing Approaches for NLG

- **Retrieval based Dialog Systems**
  - Respond to a user's turn X by repeating some appropriate turn Y from a corpus of natural (human) text
  - Given the corpus and the user's sentence, any retrieval algorithm can be adopted for this task

- **Neural Generative Dialog Systems**
  - Treat the task of NLG as the problem of transferring from one sequence to the other
  - Neural Machine Translation based approaches can be used

Recent Trends in Dialogue Systems

- **An engaging response generation system should be able to output grammatical, coherent responses that are diverse and interesting**
  - A Diversity-Promoting Objective Function for Neural Conversation Models (NAACL, 2016)
- **Emotion and Sentiment driven Dialogue Agents**
  - Desirable to understand the sentiment and emotion of the speaker while generating responses
  - Will make machines more user-friendly

    Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory (AAAI, 2018)

    SentiGAN: Generating Sentimental Texts via Mixture Adversarial Networks (IJCAI, 2018)

    Sentiment Adaptive End-to-End Dialog Systems (ACL, 2018)

    Generating Responses with a Specific Emotion in Dialog (ACL, 2019)

Some Recent Advancements in Dialogue Systems

- **Personalized dialogue agents**
  - Preserving personalized behaviour is important in today's dialogue agents
  - Training Millions of Personalized Dialogue Agents (EMNLP, 2018)
  - Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good (ACL, 2019)

- **Preserving courtesy while generating responses**
  - For NLG module (generic or task oriented), courteous response can play an important role in keeping the user engaged with the system
  - Hitesh Golchha, Mauajama Firdaus, Asif Ekbal, Pushpak Bhattacharyya; *Courteously Yours: Inducing courteous behavior in Customer Care responses using Reinforced Pointer Generator Network*. In NAACL-HLT 2019.
- **Multi-modal NLG**
  - Information from more than one modality could improve the overall performance
  - Visual Dialog (CVPR, 2017)
  - Towards Building Large Scale Multimodal Domain-Aware Conversation Systems (AAAI, 2018)

Hitesh Golchha, Mauajama Firdaus, Asif Ekbal, Pushpak Bhattacharyya (2019). *Courteously Yours: Inducing courteous behavior in Customer Care responses using Reinforced Pointer Generator Network*. In NAACL-HLT 2019, PP. 851-860

What we solve?

To **transform** a generic chatbot response into a response which uses **courteous phrases** and make the users **more engaged in conversation**

Domain: Customer Care on Twitter

Goal: To convert a generic chatbot reply into one that:

- is ***emotionally aware*** and ***intelligent***
- uses courteous phrases and emoticons to display ***appreciation***, ***empathy***, ***apology***, ***assurance***

Purpose is to increase user satisfaction and to build customer relations

Motivation and Significance

- For any NLG module (*generic* or *task oriented*), ***courteous response*** can play an important role in keeping the ***user engaged with the system***
- Makes the Chatbot more ***human-like*** while generating responses
- Inducing ***Courteous behavior*** in responses can be fused with any existing NLG system
  - to give them humanly essence and
  - simultaneously make users more comfortable in using these systems leading to an increase in user association with the brand or product
- Leads to ***customer satisfaction*** with an increase in ***customer retention*** and ***strong customer relations***

Challenges

- Unavailability of data for modelling the courteous behaviour

- Annotation of the data
  - identifying different variations and styles of courteous behaviours across different companies, service providers, demographics and cultures
  - Identifying courteous behaviours in the customer care domain is not straightforward
- Hard to model the **emotion across the conversation** for effective courteous response generation
  - System needs to capture the **correct emotion** and accordingly handle the customer by replying courteously
  - For e.g., ***if the customer is angry*, the system needs to *pacify* and *apologize***
  - **If they are *happy*, then *appreciate***

What we do here?

- Propose a very novel research direction of inducing courteous behavior in the natural language responses for the customer care domain whilst being contextually consistent

- **Create a high quality and a large-scale conversational dataset**
  - Courteously Yours Customer Care Dataset (CYCCD)
  - Prepare from the actual conversations on Twitter
  - Provide both forms of agent responses: ***generic*** and ***courteous***

- Propose a strong benchmark model based on a *context* and *emotionally aware* reinforced pointer-generator approach
  - demonstrates very strong performance (both on ***quantitative*** and ***qualitative*** analyses) on established and task-specific metrics, both automatic and human evaluation based

Courtesy: Derived from the Politeness Theory

(***A brief Introduction***)

Courtesy: The behaviour

- The showing of ***politeness*** in one's attitude and behaviour towards others
- A courtesy is a polite remark or respectful act

For example, ***complain about a bad meal,*** and you might get kicked out

BUT,

the common courtesy is usually an ***apology from the manager*** and, if you're **lucky,** a free **dinner**

Theory of Politeness

- Brown and Levinson (1987) introduced the notion of '***face***' in order to illustrate '***politeness***' in the broad sense
  - All interactants have an interest in maintaining two types of 'face' during interaction, i.e. '**positive face**' and 'negative face'
- **Positive face**
  - positive and consistent image people have of themselves, and their desire for approval
  - refers to one's self-esteem
- **Negative face**
  - the basic claim to territories, personal preserves, and rights to non-distraction
  - refers to one's freedom to act

Politeness: Dual Nature

- **Politeness**
  - Positive politeness
  - Negative politeness
- **Positive** and **negative** faces exist universally in human culture
- ***Positive politeness*** is expressed by satisfying *positive face* in two ways:
  - by indicating similarities amongst interactants; or
  - by expressing an appreciation of the interlocutor's self-image
- ***Negative politeness*** can also be expressed in two ways
  - by saving the interlocutor's ***face*** (either '*negative*' or '*positive*') by mitigating face threatening acts, such as *advice-giving* and *disapproval*; or
  - by satisfying ***negative face*** by indicating respect for the addressee's right not to be imposed on

Politeness Strategies

- **Less Polite Strategies**
  - seeking agreement
  - joking
  - expressing optimism

- More Polite Strategies
  - being apologetic
  - minimizing the imposition
  - appreciating
  - thanking

Let’s see:

Some use cases of courteous behaviours in Customer Care Systems

Use-cases of Courteousness

Generic	Courteous	Behaviour
How can we help?	*Help has arrived! We are sorry to see that you are having trouble, how can we help?*	*Apology*
Can you send us a screenshot of what you're seeing?	*Hey Craig, help's here! Can you send us a screenshot of what you're seeing?*	*Greet*
Let's discuss it in DM.	*We want to help. Let's discuss it in DM.*	*Assurance*
What is happening with your internet?	*Oh no that's not good. I can help! What is happening with your internet?*	*Empathy*
Enjoy your show while flying!	*Thanks for your kind words and enjoy your show while flying!*	*Appreciation*

Examples- ***Expressing Apology / Empathy***

somebody from @VerizonSupport please help meeeeee. I'm having the worst luck with your customer service

@115719 How can we help?

@VerizonSupport I finally got someone that helped me, thanks!

@115719 Awesome!

somebody from @VerizonSupport please help meeeeee. I'm having the worst luck with your customer service

@115719 **Help has arrived! We are sorry to see that you are having trouble.** How can we help?

@VerizonSupport I finally got someone that helped me, thanks!

@115719 Awesome! **If you ever need us we are just a tweet away.**

Example- ***General Assistance***

@SpotifyCares Is the Radio X radio within the app a @320735 branded station, or just a station called Radio X?

"@320734 Can you send us a screenshot of what you're seeing? We'd like to take a close look",

@SpotifyCares https://t.co/EGzHYLFXDf

@320734 This playlist radio created based on the playlist Radio X you played.

@SpotifyCares Thanks

"@320734 Happy listening ",

@SpotifyCares Is the Radio X radio within the app a @320735 branded station, or just a station called Radio X?

"@320734 ***Hey Craig, help's here!*** Can you send us a screenshot of what you're seeing? We'd like to take a close look",

@SpotifyCares https://t.co/EGzHYLFXDf

@320734 ***Thanks!*** This playlist radio created based on the playlist Radio X you played.

@SpotifyCares Thanks

"@320734 ***No worries at all! For anything else, we're just a tweet away.*** Happy listening 🙂",

Example-***Expressing Assurance***

so my iPhone was perfectly working a minute back and then it stopped working just out of nowhere. what kind of a sorcery is this

@294072 What trouble are you having?

my iphone's screen stopped responding in the middle of nowhere as if it got busy. its been 15 long hours yet NO hope. how come?

@294072 Let's discuss it in DM. https://t.co/GDrqU22YpT

so my iPhone was perfectly working a minute back and then it stopped working just out of nowhere. what kind of a sorcery is this

@294072 What trouble are you having? ***We'll look into this with you and see what's going***

my iphone's screen stopped responding in the middle of nowhere as if it got busy. its been 15 long hours yet NO hope. how come?

@294072 ***We want to help.*** Let's discuss it in DM. https://t.co/GDrqU22YpT

Example-*Expressing Empathy and Assurance*

@CoxHelp y'all just came to my house like last week and I'm having problems with my internet again smh

@294080 What is happening with your internet?

@CoxHelp I think there's an outage

@294080 There may be maintenance work in your area. Please DM your info.

@CoxHelp y'all just came to my house like last week and I'm having problems with my internet again smh

@294080 **Oh no that's not good. I can help!** What is happening with your internet?

@CoxHelp I think there's an outage

@294080 There may be maintenance work in your area. Please DM your info. **Thanks for using our services.**

Example- **Expressing Appreciation**

Thank you, @AmericanAir for playing #ThisIsUs and for having great flight attendants on my flight back home!

@115909 We're glad you got to kick back and enjoy a show while flying!

Thank you, @AmericanAir for playing #ThisIsUs and for having great flight attendants on my flight back home!

@115909 We're glad you got to kick back and enjoy a show while flying! **Thanks for your kind words.**

Resource Creation: Data source and Attributes

- **Twitter dataset from Kaggle**
  - Interactions between customers and professional customer care agents of companies
  - Tweets have company names, anonymized user ids, time stamps, and response tweet ids

- **Pre-processing**
  - Segment the tweet into sentences
  - Remove purely ***courteous (and non-informative)*** sentences
  - Retain purely informative sentences
  - Transform the informative sentences with courteous expressions (***to remove only the courteous part from the sentence***)

Pre-processing: Three-form of Sentences (contd..)

- **Courteous sentences**
  - Sentences which do not contain any information or suggestion and purely non-informative
  - Include personalized greetings and expressions of appreciation, apology, empathy, assurance, enthusiasm
  - Example: *Sorry to hear about the trouble*

- **Informative sentences without courteous expressions**
  - Contain actual contents of the tweet and are generally assertions, instructions, imperatives or suggestions
  - Example: *Simply visit url_name to see availability in that area*

- **Hybrid-*informative sentences with courteous expressions***
  - Combination of above two
  - Example: *We appreciate the feedback, we'll pass this along to the appropriate token*

Resource Creation: Scaling up for large data creation

- **Clustering**
  - Expressions and sentences used by company professionals are of similar patterns
  - Obtain the vector-semantic representations of sentences using the sentence encoder trained on the SNLI corpus
  - k-means clustering (k = 300) used to cluster these sentences
- **Annotations**
  - Three annotators proficient in English language were assigned to annotate the sentences into the three categories
    - purely courteous
    - purely informative
    - hybrid
  - Multi-rater Kappa agreement ratio: **85.18%**

Resource Creation: Scaling up for large data creation

- **Preparing generic responses**
  - Obtain the generic response by removing

    ***Courteous sentences*,**

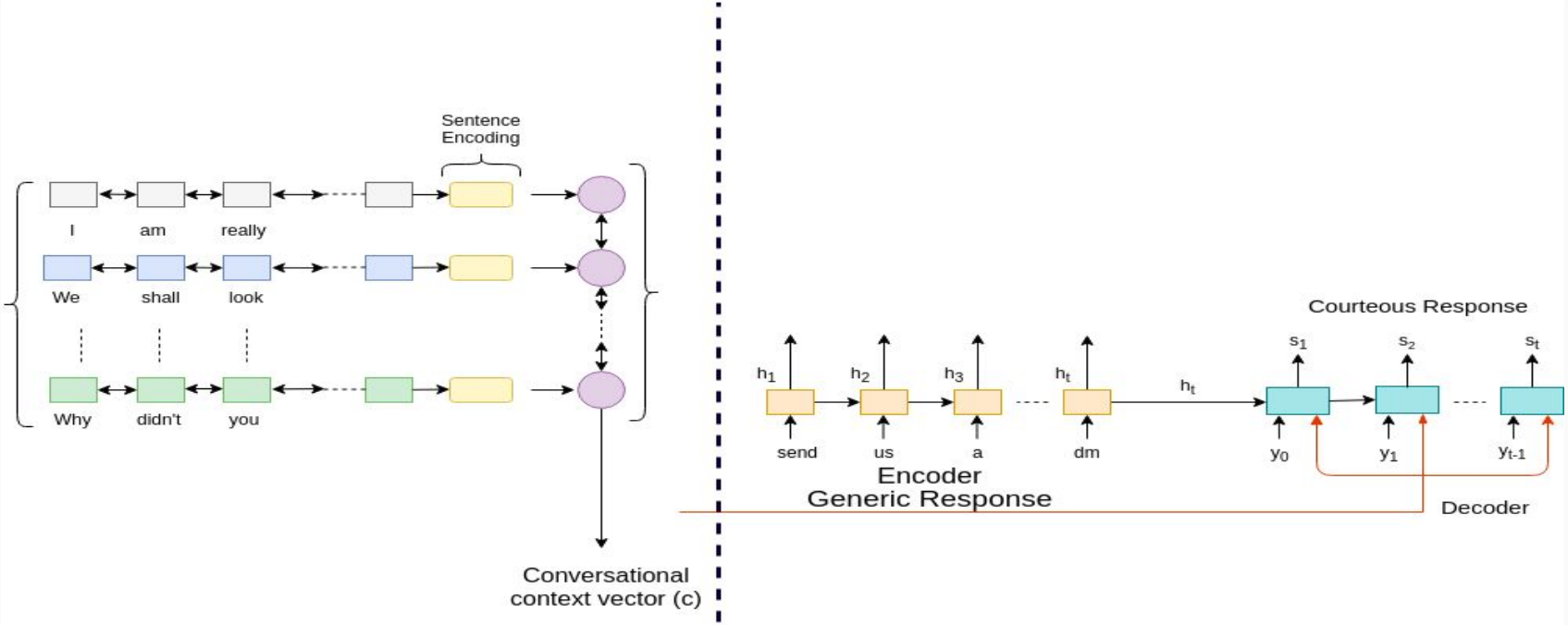
    retaining the informative sentences, and

    replacing the hybrid sentences with the prepared generic equivalents

Dataset: Statistics

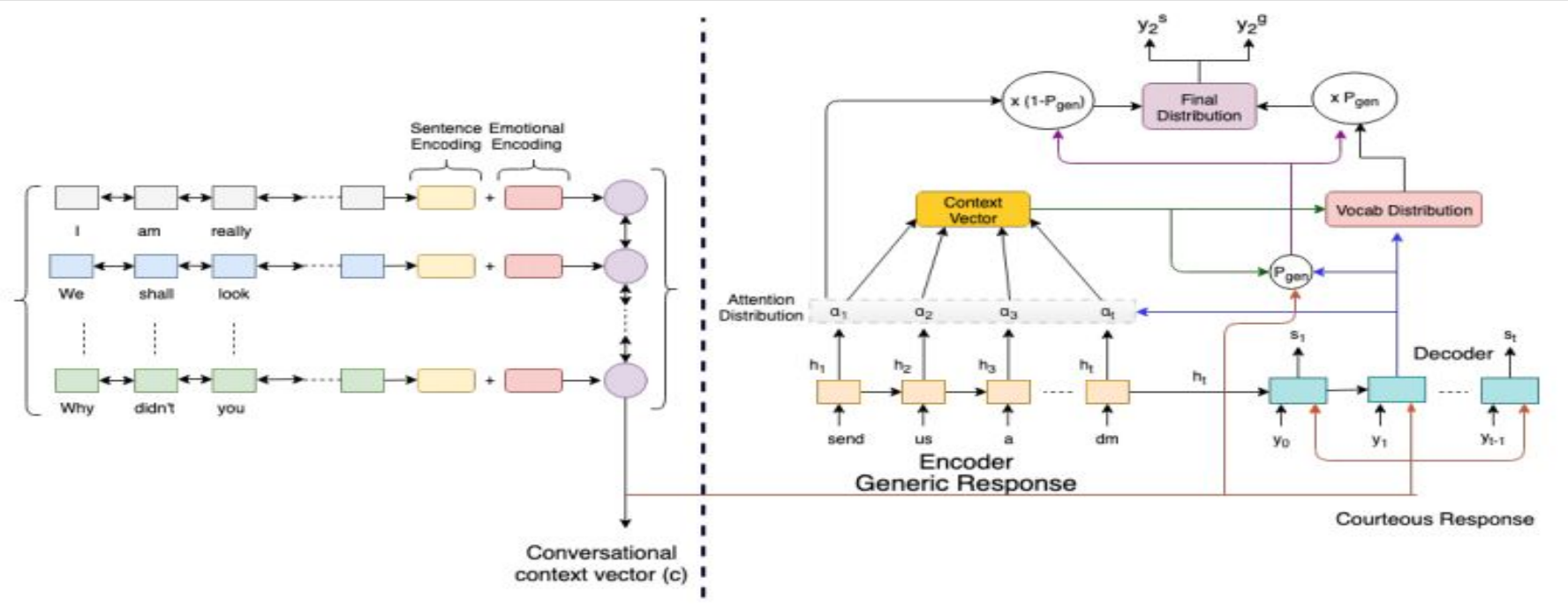
	Train	Validation	Test
# Conversations	140203	20032	40065
#Utterances	179034	25642	51238

Baseline Model



Input to the Model: Generic Response and Conversational History, Output: Courteous Response

Proposed Methodology



Inputs to the model: Conversation History (left), Generic Response (centre) Output: Courteous Response (right). The Conversation History is encoded by hierarchical Bi-LSTM to a Conversational Context vector *c*. The encoder encodes the Generic Response into hidden states h_i . Response tokens are decoded one at a time. Attention a_i, and vocabulary distributions (p_{vocab}) are computed, and combined using p_{gen} to produce output distribution. Sampling it yields y^s_i and taking its argmax yields y^g_i

Model Training

- Use the joint reinforcement learning (RL) and machine learning (ML) training
- Maximum likelihood objective given by:

$$L_{MLE} = -\sum_{t=1}^{n'} \log p(\tilde{y}_t | \tilde{y}_1, \ldots, \tilde{y}_{t-1}, x_1, x_2)$$

Where, y_i is the gold output token, x_1 is the generic response, x_2 is the conversational history, n' is the number of tokens in the sequence

- Along with MLE, reinforcement learning is used to learn from maximizing discrete metrics that are task-specific
- During training, two output sequences are produced: y^s, obtained by random sampling $p(y^s_t \mid y^s_1, \ldots, y^s_{t-1}, x)$ probability distribution, and y^g, the ***baseline output***, obtained by greedily maximizing the output probability distribution at each time step

Model Training

- For RL training, the probability distribution is given by:

$$L_{RL} = (r(y^g) - r(y^s)) \sum_{t=1}^{n'} \log p(y_t^s | y_1^s, \ldots, y_{t-1}^s, x_1, x_2)$$

  Where, y_i are the gold output tokens, x_1 is the generic response, x_2 is the conversational history
- **Reward function r_y**, used for evaluating y against the gold standard output
  - Weighted combination of the following two metrics
    - **BLEU metric:** Ensures the content matching between the reference and the decoded outputs
    - **Emotional accuracy**
      - Measured by the cosine similarity of the emoji distributions of the gold and generated responses (***using pre-trained DeepMoji***)
      - Ensures that the emotional states of the generated courteous behavior is consistent with the gold standard

Model Training

- Our reward function r(y), used for evaluating y against the gold standard output:

$$r(y, \tilde{y}) = \lambda_1 \cdot m1(y, \tilde{y}) + \lambda_2 \cdot m2(y, \tilde{y})$$

Where, m1 : BLEU metric, m2: Emotion accuracy, λ1 and λ2 are 0.75 and 0.25, respectively

- If Reward difference = positive, i.e., the BLEU score of $r(y^g) > r(y^s)$
  Then, the log probability of the generated response by random sampling will be decreased

- If Reward difference = negative, i.e., the BLEU score of $r(y^g) < r(y^s)$
  Then, the log probability of the generated response by random sampling will be increased

- We first pre-train using the maximum likelihood (MLE) objective, and then using a mixed objective function with a reduced learning rate (η):

$$L_{mixed} = \eta L_{RL} + (1 - \eta) L_{MLE}$$

Evaluation Metrics, Results and Analysis

Evaluation Metrics: Automatic

- ***Content Preservation***
  - Measures how much of the informative content from the original generic response is reflected in the generated courteous response
- ***Emotional Accuracy***
  - Measures the consonance between the generated courteous expressions (source of emotion) and the gold
  - Determine the similarity between the emoji distributions of the two responses

Evaluation Metrics: Human

- ***Fluency***: Measure whether courteous response is grammatically correct and is free of any errors
- ***Content Adequacy***: Generated response contains the information present in the generic form of the response and there is no loss of information while adding the courteous part to the responses
- ***Courtesy Appropriateness*****:** The courtesy part added to the generic responses is in accordance to the conversation history

  - Scoring scheme for fluency and content adequacy:
    0-incorrect or incomplete; 1: moderately correct; 2: correct
  - Scoring scheme for courtesy appropriateness
    -1: in-appropriate; 0-non-courteous; 1: appropriate

Baselines

- Model-1 (Luong et al., 2015)
  - Seq2Seq model with attention
  - Decoder conditioned on the conversational context vector
  - No emotional embeddings

- Model-2
  - Model-1 +
  - copying mechanism of pointer generator network

- Model-3
  - Model-2 +
  - Emotional embeddings

Evaluation Results- Automatic

	Model	BLEU	ROUGE			PPL	CP	EA
			1	2	L			
1	*Seq2Seq*	56.80	63.8	59.06	64.52	58.21	68.34	82.43
2	*Seq2Seq + P*	66.11	69.92	64.85	66.40	**42.91**	**77.67**	81.98
3	*Seq2Seq + P + EE*	68.16	72.18	67.92	71.17	43.52	76.05	85.75
4	*Proposed Model*	**69.22**	**73.56**	**69.92**	**72.37**	43.77	77.56	**86.87**

P: Pointer Generator Model; EE: Emotional embedding; PPL: Perplexity; CP: Content Preservation; EA: Emotion Accuracy

- **Observations**:
  - Model-2 is aided by copying mechanism, and hence performance is improved as it can copy portions from generic response and forward to courtesy
  - Model-3 improves the performance by 3.77 (EA) over Model-2 as it can better understand the emotional states and generate more courteous responses
  - Perplexities in Model-3 and Model-4 are more compared to Model-2 : may be due to emotional embeddings that confuse generated response from the ground truth

Evaluation Results- Human

Model	F			CA			CoA		
	0	1	2	0	1	2	-1	0	1
Model 1	15.70	42.50	41.80	16.21	41.69	42.10	23.71	51.08	25.21
Model 2	14.23	42.77	43.00	15.62	39.65	44.73	22.05	39.43	38.52
Model 3	11.15	44.10	44.75	13.66	41.12	45.22	15.23	41.22	43.55
Our Model	10.05	44.90	44.60	13.85	38.48	47.67	14.11	41.11	44.78

F: Fluency, CA: Content Adequacy and CoA: Courtesy Appropriateness

Some example outputs

Generic Input	Model 1	Model 2	Model 3	Our Model
dm us more info and well take a look into it for you	we'll look into it	im sorry to hear this please dm us more info and we'll take a look into it for you	were here to help please dm us more info and well take a look into it for you	were here to help please dm us more info and well take a look into it for you at the earliest
adjust the brightness via your display settings on your device	whos the brightness via your display settings on your device	were here to help adjust the brightness via your display settings on your device	we have several ways to change the display brightness on your device and were happy to help	thanks for reaching out we have several ways to change the display brightness on your device and were happy to help
we'll follow up with the store	we'd like to help well follow up	were here to help well follow up with the store	sorry to hear that well follow up with the store	thats disappointing to hear, we'll follow up with the store
can you confirm which platform you are using for video access ? what is the error ?	what is the error ?	I am sorry for the frustration ! can you confirm which platform you are using for video access ? what is the error ?	I am sorry to hear this can you confirm which platform you are using for video access? what is the error?	I am sorry for any frustration, can you please confirm which platform you are using for video access? Please tell us what is the error.
fill this form <url>	please fill this form <url>	were here to help fill this form <url>and I'll contact you at the earliest a	apologies for the hassle, please fill this form <url>and we'll contact you thank you for reaching out to us we will follow up with the store	i am sorry for the hassle, please fill this form <url> and ill contact you at the earliest

Error Analysis

- Unknown Tokens
  - Model 1 suffers the most as it does not have any copying mechanism
  - Often the model predicts ‘end of sequence’ token just after the ‘out of vocabulary’ token, thus leaving sequences incomplete

- Wrong Copying
  - Sometimes pointer network makes mistakes while copying (being influenced by language model)

    Example: Gold: .. which store in gillingham did you visit ?; Predicted: .. which store in belgium did you visit ?

- Errors in Emotion Identification
  - More prominent in Models 1 and 2 (they don’t have emotional embeddings), where the generated courteous phrases denote mistakes in identifying the emotional state of the customer

Example,
Gold: you’re very welcome, hope the kids have an amazing halloween !; Predicted: we apologize for the inconvenience. hope the kids have an amazing halloween !

Error Analysis

- Extra information
  - Models 1, 2, 3 sometime generate extra informative sentences than in the generic response
    Example:
    Gold: please send us a dm; Predicted: please send us a dm please let us know if you did not receive it

- Contextually wrong courteous phrases
  - These mistakes are common across models while generating courteous phrases with content in them
    Example:
    Gold: we want to help, reply by dm and ..; Predicted: im sorry you haven't received it. Please reply by dm and ..

- Difference in phrases
  - Generated responses differ from reference responses in their use of (equivalent) courteous phrases, and are hence wrongly penalized by some metrics

Hardik Chauhan, Mauajama Firdaus, Asif Ekbal, Pushpak Bhattacharyya (2019). Ordinal and Attribute Aware Response Generation in a Multimodal Dialogue System. ACL 2019: 5437-5447

The problem we solve

- NLG in Multi-modal Dialogue Systems
- **Overall System**
  - Input: Text and Image
  - Output: Appropriate response

- Dialogue consists of text utterances along with multiple images
  - Given a context of k turns the task here is to generate the next text response

- Addressed the task of textual response generation conditioned on conversational history
- Introduced the novel idea of incorporating the **position** and **attribute-aware attention mechanism**

Background: Why Multimodality Important?

- Majority of the existing research on dialogues concentrate only on text
- Phenomenal growth of information through web, e-commerce sites, social networking sites in terms of audios, videos, images along with text
- Conversational AI has great potential for online retail, travel, entertainment
  - Greatly enhances user experience and
  - Directly affects user retention
- Knowledge from different modalities carries complementary information about the various aspects of a product, event or activity of interest
  - Effective combination of the information from different modalities is crucial for creating robust dialogue systems

Motivating Example- I/2 (*Showing importance of position and attributes of image*)

- **Position information is important**
  - 7th Utterance (Fifth image)
- Incorporate position-aware embeddings to capture ordered visual information
  - To capture the correct image information from text

- **Different attributes of the product are important**
  - 3rd utterance (*blue* color)
  - 7th Utterance (*material* and *style*)
- Incorporate attribute-aware attention
  - To focus and capture different attributes from the image representations

SHOPPER: Hello

AGENT: Hello, is there anything i can help you with today?

SHOPPER: show me some espadrilles in beige/blue within $200.

AGENT: Sure. let me just quickly browse through my catalogue

AGENT:

AGENT: What you can see in the fifth item, michael michael kors brings you an effortless weekend look with the wrapped espadrille heel and crisscross styling of these dana flats. And what you see in the 3rd item, steve madden, the company, was founded by steve madden, the man, in 1990 in brooklyn, where he started out making trend-advancing shoes one pair at a time.

SHOPPER: List more in the upper material as in the 5th image and style as the 3rd and the 5th

AGENT:

AGENT: The 2nd one, viscata espadrilles are 100% authentic and handcrafted in spain inspired by the mediterranean, viscata shoes are lightweight and made of natural jute rope, premium leathers, suedes and woven cloth.

SHOPPER: Can you list images of the 1st result from all orientations?

AGENT: Image from the front,right and back orientations respectively

SHOPPER: And the 3rd one?

AGENT: Image from the front,right,back and left orientations respectively

Motivating Example- 2/2 (*Showing importance of position and attributes of image*)

- **Position is important**
  - 11th utterance (***1st image***)
- **Attribute is also important**
  - 3rd utterance (***formal***)
  - 11th utterance (***1st image***)

SHOPPER: Hello

AGENT: Hi, please tell me what i can help you with?

SHOPPER: I need to shop formal trouser sporting a loop closure for myself .

AGENT: Can you please tell me your gender?

SHOPPER: male

SHOPPER: I am a 16 year old guy.

AGENT: Okay Good!

AGENT: let me just quickly search through my catalogue of items

AGENT: 

AGENT: What you can see in the second item, take a cue from stylish gents around the globe and build your wardrobe with polished and versatile pieces, such as our linen trousers.

SHOPPER: I dont like the style in the 1st image. Can you show me some more

AGENT: 

AGENT: About the fourth one, get ready to complete your everyday business looks this season with these navy blue coloured trousers from jack & jones. See the 5th one, flaunt the youthful, laidback look wearing beige coloured slim fit formal trousers by jctex. And to describe the third image, get complimented upon for your spectacular dressing sense by wearing these well-tailored formal trousers for men by jack & jones.

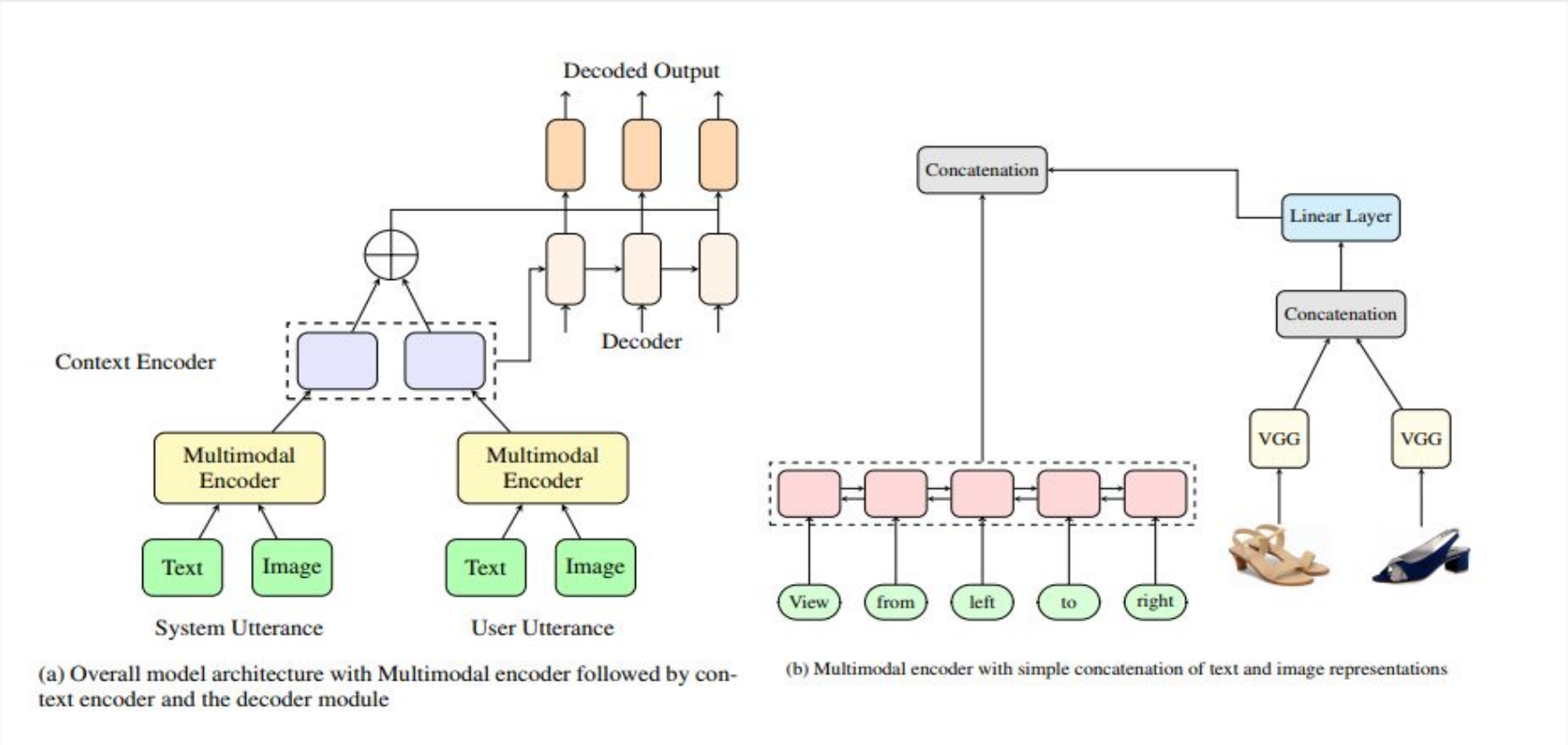
SHOPPER: Sort these by best seller ranking

AGENT: Sorry i do not have the best seller ranking information for these

SHOPPER: Thank you for showing me all this, i guess i will leave

AGENT: Sure, would be glad if you can provide a feedback on your experience

Proposed Methodology: Baseline Model



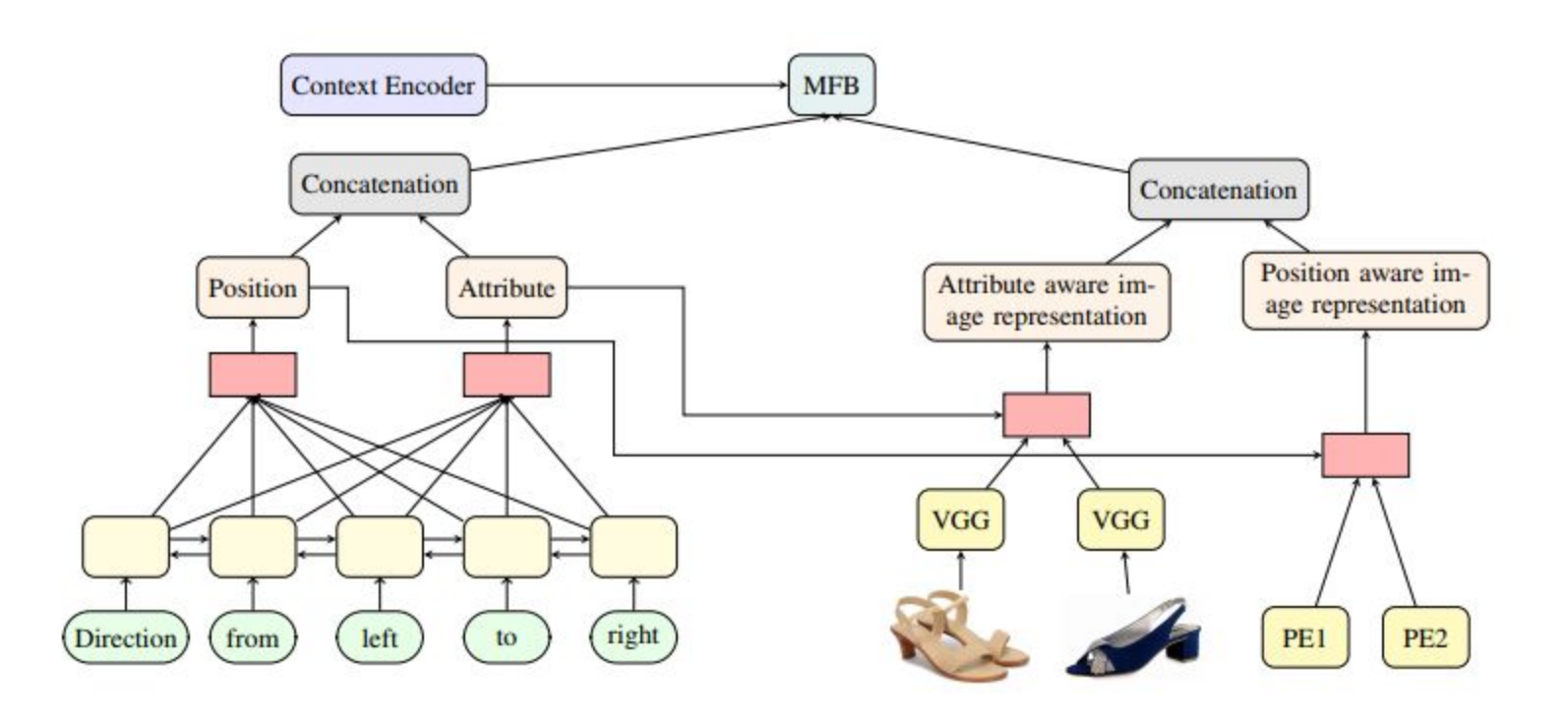
Left image is the overall system architecture for text generation; Right image is the baseline encoder model
Utterance Encoder: Bi-GRU; **Image Encoder**: VGG-19; **Global image vector**: concatenated vector of all images

Limitations of the Existing Baseline Multimodal Encoder

- Does not have the capability of incorporating position and attribute related information of images
  - Revealed from the examples that position and attribute of images are essential for the system to fulfill the demands of the user

- No interaction between the modalities and the features are merely concatenated for generating the responses
  - For effective interaction among the modalities, we use Multimodal Factorized Bilinear (MFB) pooling mechanism (Yu et al., 2017)
  - Information of the present utterance, image and the contextual history are essential for better response generation (Serban et al., 2015)

Proposed Methodology



Proposed Multimodal Encoder with Position and Attribute aware Attention with MFB fusion

Dataset

- Proposed Multimodal Dialogue (MMD) dataset (Saha et. al.)
  - 150K Chat conversations

	Train	Validation	Test
Number of dialogues	105,439	22,595	22,595
Avg. turns per dialogue	40	40	40
No. of utterances with image response	904K	194K	193K
No. of utterances with text response	1.54M	331K	330K
Avg. no words in text response	14	14	14

Evaluation Metrics

- **Automatic Evaluation Metrics**
  - BLEU-4 (Papineni et al., 2002)
  - METEOR (Lavie amd Agarwal, 2007)
  - ROUGE-L (Lin, 2004)

- **Human Evaluation Metrics**
  - **Fluency**: Generated response is grammatically correct and is free of any error
  - **Relevance**: Generated response is in accordance to the aspect being discussed (style, colour, material, etc.), and contains the information with respect to the conversational history
  - Scoring scheme: 0-***incorrect or incomplete***; 1-***moderately correct***; 2: ***correct***

Evaluation Results: Automatic

Description	Model	BLEU 4	METEOR	ROUGE L
State-of -the-arts	*MHRED-attn (Agarwal et al., 2018a)*	0.4451	0.3371	0.6799
	MHRED-attn-kb (Agarwal et al., 2018b)	0.4634	0.3480	0.6923
Baseline Models	*MHRED*	0.4454	0.3367	0.6725
	MHRED + A	0.4512	0.3452	0.6754
	MHRED + A + PE	0.4548	0.3476	0.6783
	MHRED + PA	0.4781	0.3521	0.7055
	MHRED + AA	0.4763	0.3511	0.7063
	MHRED + PA + AA	0.4810	0.3569	0.7123
	MHRED + MFB(I,T)	0.4791	0.3523	0.7115
	MHRED + MFB(I,T,C)	0.4836	0.3575	0.7167
Our Proposed Model	***MHRED + PA + AA + MFB(I,T)***	**0.4928**	**0.3689**	**0.7211**
	MHRED + PA + AA + MFB(I,T,C)	**0.4957**	**0.3714**	**0.7254**

MHRED: Multi-modal Hierarchical Encoder Decoder, A: Attention, PE: Positional embeddings, PA:Position-aware attention, AA: Attribute-aware attention, MFB (I,T): MFB fusion on image (I) and text (T) representations, MFB(I,T,C): MFB fusion on I,T and context (C)

Evaluation Results- Human

Description	Model	Fluency			Relevance		
		0	1	2	0	1	2
Baseline	*MHRED*	18.64	39.66	41.70	13.41	39.83	46.76
Proposed	*MHRED + PA + AA + MFB(I,T,C)*	15.54	42.71	41.75	7.36	38.14	54.23

Observations:

- For **fluency**, MHRED (baseline) and proposed model exhibit similar performance
- For **relevance**, proposed model performs superior (with 7.47% improvement)
  - May be due to the efficacy of our model to focus on the relevant information in the text as well as the image, and generate more accurate and informative responses

Evaluation: Attention Visualization

Example 1:

USER: I like the weave in the 3rd one but not the type. Can you show me some more?

Example 2:

USER: I liked the 2nd high tops. Can I see something like it but containing the sole made out of rubber material.

Example 3:

USER: I like the vintage wayfarer style sunglasses but in dark lenses and red frame.

Observations:

- Example-1: Model can focus on the correct image (here, the 3rd image)
- Example-2: Shows the effect of both position and attribute aware attention mechanism (position- 2nd; Attribute: rubber)
- Example-3: Effect of attribute-aware attention is evident (with more focus on the keywords such as *dark*, *red*, *frame*)

Output responses generated by different models: Few examples

Example 1:

USER: What is the color and style of the first result?

Ground Truth: The shoe in the first image has black formal type

Baseline: The shoe in the first image has grey casual type

Proposed: The shoe in the first image has black formal type

Example 2:

USER: Can you tell me the style of the 4th result?

Ground Truth: The style of the 4th is contemporary and classic

Baseline: The shoe in the first image has grey casual type

Proposed: The style of the 4th image is classic

Example 3:

USER: Can you show hand block printed pink anarkali kurti?

Ground Truth: We have mustard colored suit set tailored from polyster in a flared kurti style.

Baseline: We have mustard colored kurti.

Proposed: Mustard colored, polyster material in flared style kurti available.

Sentiment and Emotion Controlled Multimodal Dialogue Generation

Problem definition and Motivation

To make a conversational AI agent enabled to generate sentiment and emotion controlled dialogue

Input: Text, Audio and Video

Output: Sentiment and Emotion aware response based on conversational history

- **Motivation**
  - To make an intelligent agent capable of understanding and generating responses according to human emotions and sentiments (***one of the greatest challenges of any AI system***)
  - Eventually leading to user satisfaction and customer attention

An Example from SEMD Dataset

1) The man who passes the sentence should swing the sentence.

Anger
(**Negative**)

2) Is it true he saw the white walkers?

Surprise
(**Negative**)

3) The white walkers have been gone for thousands of years.

Sadness
(**Negative**)

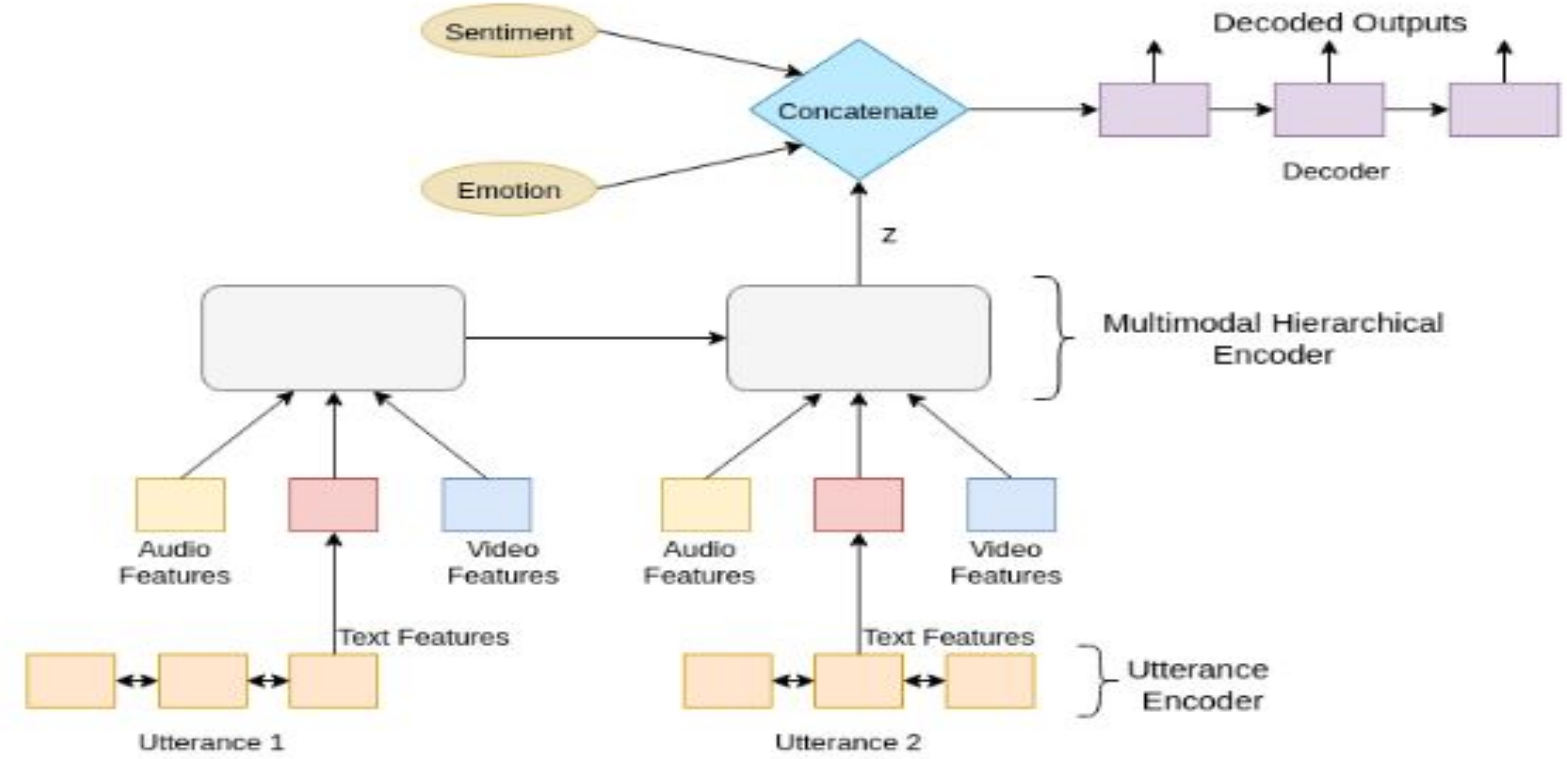
4) So he was lying?

Disgust
(**Negative**)

5) A madman sees what he sees.

Disgust
(**Negative**)

Architectural Diagram



Experimental Results

Model Description	Parameter	Modality	PPL	distinct-1	distinct-2	SA	EA
Baseline MHRED	*Sentiment (S)*	*T*	26.4	0.0065	0.017	0.65	-
		T + A	26.2	0.0067	0.019	0.68	-
		T + A + V	25.9	0.0068	0.020	0.71	-
	Emotion (E)	*T*	24.8	0.0062	0.022	-	0.62
		T + A	24.5	0.0064	0.025	-	0.63
		T + A + V	24.1	0.0066	0.026	-	0.66
	S + E	*T*	23.5	0.0078	0.036	0.73	0.70
		T + A	23.1	0.0079	0.038	0.76	0.71
		T + A + V	22.8	0.0081	0.039	0.77	0.72
Proposed CVAE Model	*S + E*	*T + A + V*	18.1	0.017	0.045	0.83	0.80

Summary and Conclusions

- **Presented a very novel study for courtesy response generation in a Conversational AI**
  - Very useful for many sectors such as retails, customer care centres etc.
  - Can retain the customers with politeness behaviors

- **Presented a multimodal NLG system for Fashion domain**
  - Modalities - text + Image
  - Attribute-aware attention
  - Position-aware attention

- **Presented Emotion and Sentiment controlled dialogue generation**
  - One of the challenges that current conversational AI agent should be dealt with

Selected Publications

1. Mauajama Firdaus, Shobhit Bhatnagar, Asif Ekbal, Pushpak Bhattacharyya; I*ntent Detection for Spoken Language Understanding Using a Deep Ensemble Model*; In proceedings of 15th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2018); 629-642; Nanjing, China; 2018.

2. Mauajama Firdaus, Shobhit Bhatnagar, Asif Ekbal, Pushpak Bhattacharyya; *A Deep Learning based Multi-task Ensemble Model for Intent Detection and Slot Filling in Spoken Language Understanding*. In proceedings of 25th International Conference on Neural Information Processing (ICONIP 2018); 647-658; Siem Reap, Cambodia; 2018.

3. Hitesh Golchha, Mauajama Firdaus, Asif Ekbal, Pushpak Bhattacharyya; *Courteously Yours: Inducing courteous behavior in Customer Care responses using Reinforced Pointer Generator Network*; In proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019); Minneapolis, USA; 2019.

4. Hardik Chauhan, Mauajama Firdaus, Asif Ekbal, Pushpak Bhattacharyya; *Ordinal and Attribute Aware Response Generation in a Multimodal Dialogue System*. In proceedings of 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019) ; Florence, Italy; 2019.

5. Mauajama Firdaus, Ankit Kumar, Asif Ekbal, Pushpak Bhattacharyya; *A Multi-Task Hierarchical Approach for Intent Detection and Slot Filling*; Knowledge Based Systems (KBS), Elsevier, 2019.

References

1. A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra, “Visual dialog,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 326–335, 2017.

2. P. J. Price, “Evaluation of spoken language systems: The atis domain,” in Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990, 1990.

3. B. Liu and I. Lane, “Joint online spoken language understanding and language modeling with recurrent neural networks,” arXiv preprint arXiv:1609.01462, 2016.

4. B. Liu and I. Lane, “Attention-based recurrent neural network models for joint intent detection and slot filling,” arXiv preprint arXiv:1609.01454, 2016.

5. D. Hakkani-Tür, G. Tür, A. Celikyilmaz, Y.-N. Chen, J. Gao, L. Deng, and Y.-Y. Wang, “Multi-domain joint semantic frame parsing using bi-directional rnn-lstm.,” in INTERSPEECH, pp. 715–719, 2016.

6. X. Zhang and H. Wang, “A joint model of intent determination and slot filling for spoken language understanding.,” in IJCAI, pp. 2993–2999, 2016.

7. Y. Wang, Y. Shen, and H. Jin, “A bi-model based rnn semantic frame parsing model for intent detection and slot filling,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), vol. 2, pp. 309–314, 2018.

References

8. C.-W. Goo, G. Gao, Y.-K. Hsu, C.-L. Huo, T.-C. Chen, K.-W. Hsu, and Y.-N. Chen, “Slot-gated modeling for joint slot filling and intent prediction,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), vol. 2, pp. 753–757, 2018.

9. O. Vinyals and Q. Le, “A neural conversational model,” arXiv preprint arXiv:1506.05869, 2015.

10. X. Wu, A. Martinez, and M. Klyen, “Dialog generation using multi-turn reasoning neural networks,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), vol. 1, pp. 2049–2059, 2018.

11. D. Raghu, N. Gupta, et al., “Hierarchical pointer memory network for task oriented dialogue,” arXiv preprint arXiv:1805.01216, 2018.

12. X. Zhou and W. Y. Wang, “Mojitalk: Generating emotional responses at scale,” in ACL, 2018.

13. C. Huang, O. Zaiane, A. Trabelsi, and N. Dziri, “Automatic dialogue generation with expressed emotions,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), vol. 2, pp. 49–54, 2018.

References

14. T. Niu and M. Bansal, “Polite dialogue generation without parallel data,” TACL, vol. 6, pp. 373–389, 2018.

15. A. Saha, M. M. Khapra, and K. Sankaranarayanan, “Towards building large scale multimodal domain-aware conversation systems,” in Thirty-Second AAAI Conference on Artificial Intelligence,pages 696-704., 2018.

16. S. Agarwal, O. Dusek, I. Konstas, and V. Rieser, “A knowledge-grounded multimodal search-based conversational agent,” arXiv preprint arXiv:1810.11954, 2018.

17. S. Agarwal, O. Dusek, I. Konstas, and V. Rieser, “Improving context modelling in multimodal dialogue generation,” arXiv preprint arXiv:1810.11955, 2018.

18. Z. Yu, J. Yu, J. Fan, and D. Tao, “Multi-modal factorized bilinear pooling with co-attention learning for visual question answering,” in Proceedings of the IEEE International Conference on Computer Vision, pp. 1821–1830, 2017.

19. R. Paulus, C. Xiong, and R. Socher, “A deep reinforced model for abstractive summarization,” CoRR, vol. abs/1705.04304, 2017.

Thank you for your attention!