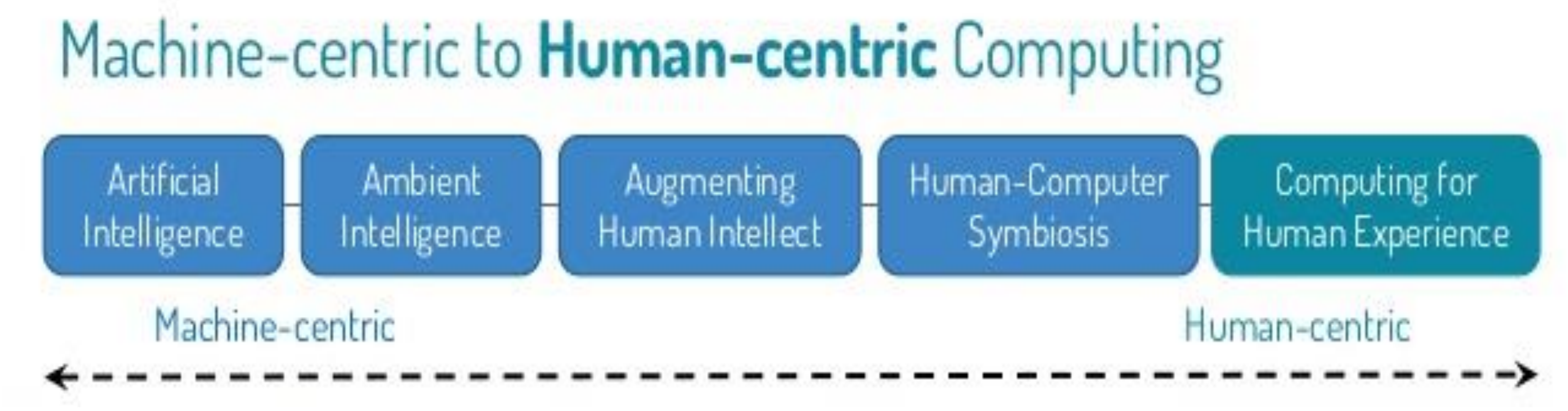
Master Thesis Proposal Talk

Towards Building A Human-like Emotional Conversational Agent Using Neural Networks

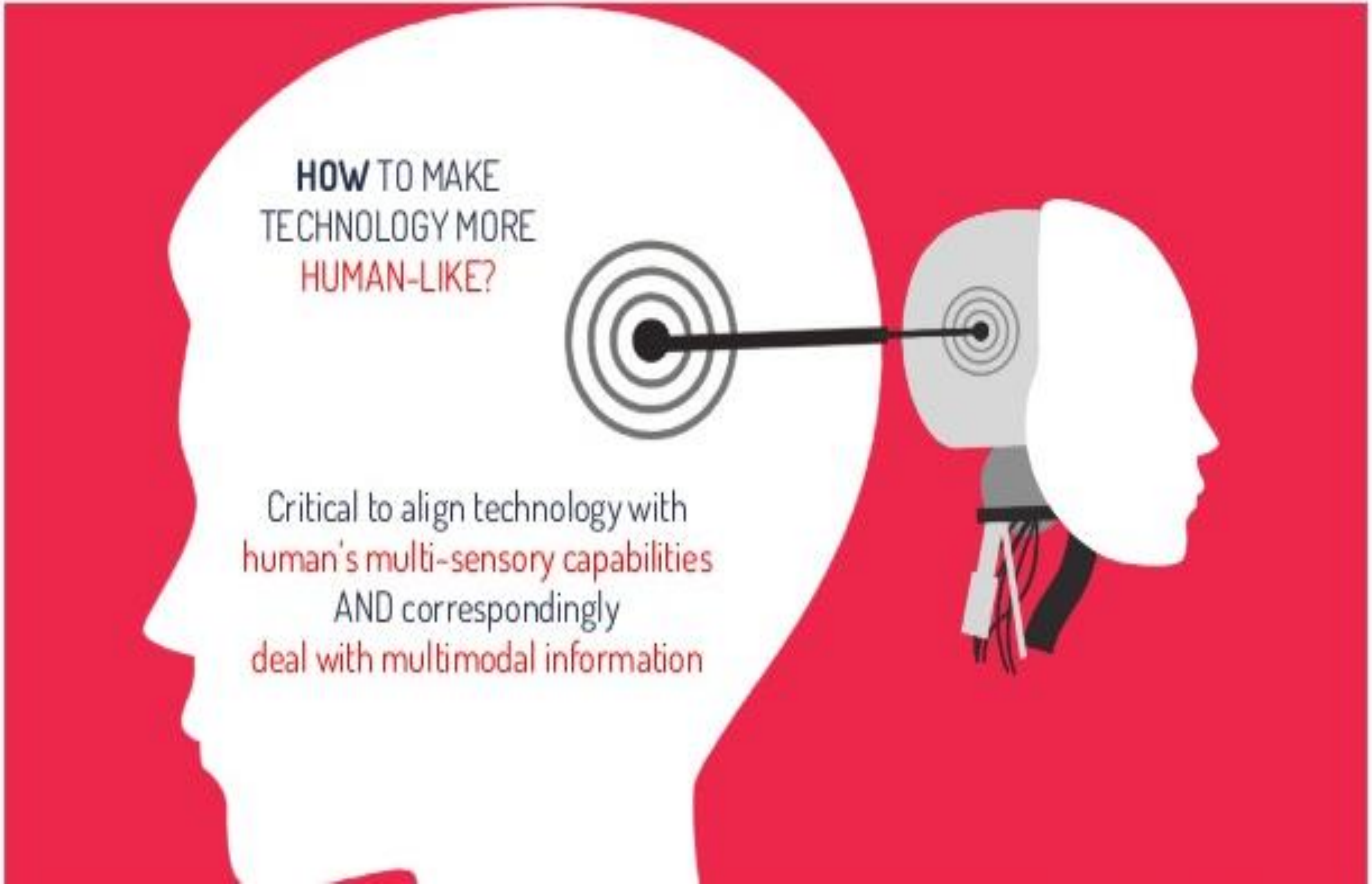
Harshita Jhavar
Advisor: Prof. Dr. Dietrich Klakow

Source: https://www.pardot.com/blog/introducing-emoji/



My Thesis

Source: https://www.slideshare.net/apsheth/humanlike-chatbots-promises-challenges-and-implications-102794389



Source: https://www.slideshare.net/apsheth/humanlike-chatbots-promises-challenges-and-implications-102794389

INSURANCE Chatbot Ask ABIE

Allstate Business Insurance expert (or ABIe)

https://hbr.org/2016/07/how-companies-are-benefiting-from-lite-artificial-intelligence

FINANCIAL Chatbot eno

Capital One Financial (Eno)

Info on credit card balances, transactions, due dates, and limit

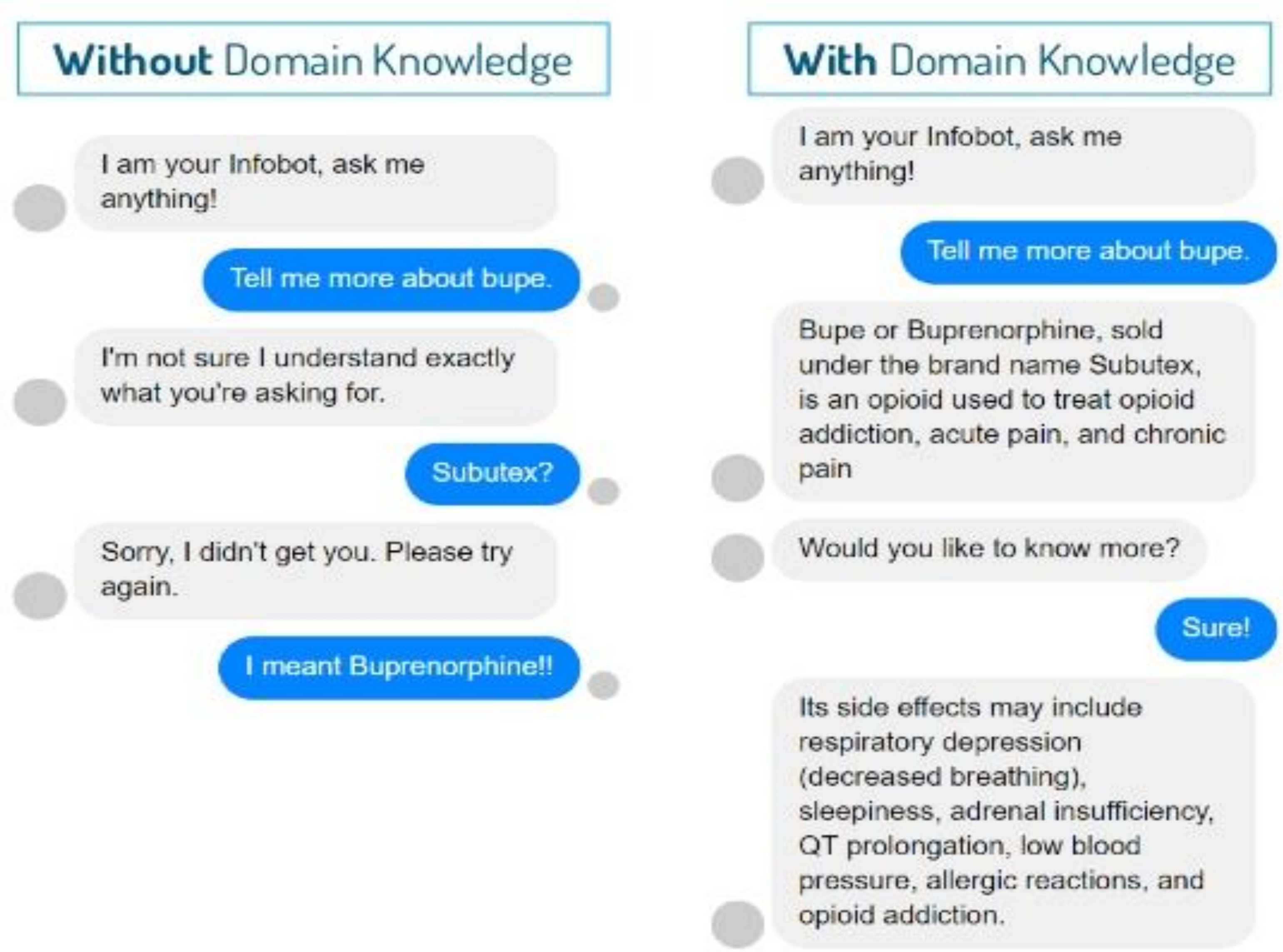
https://www.capitalone.com/applications/eno/

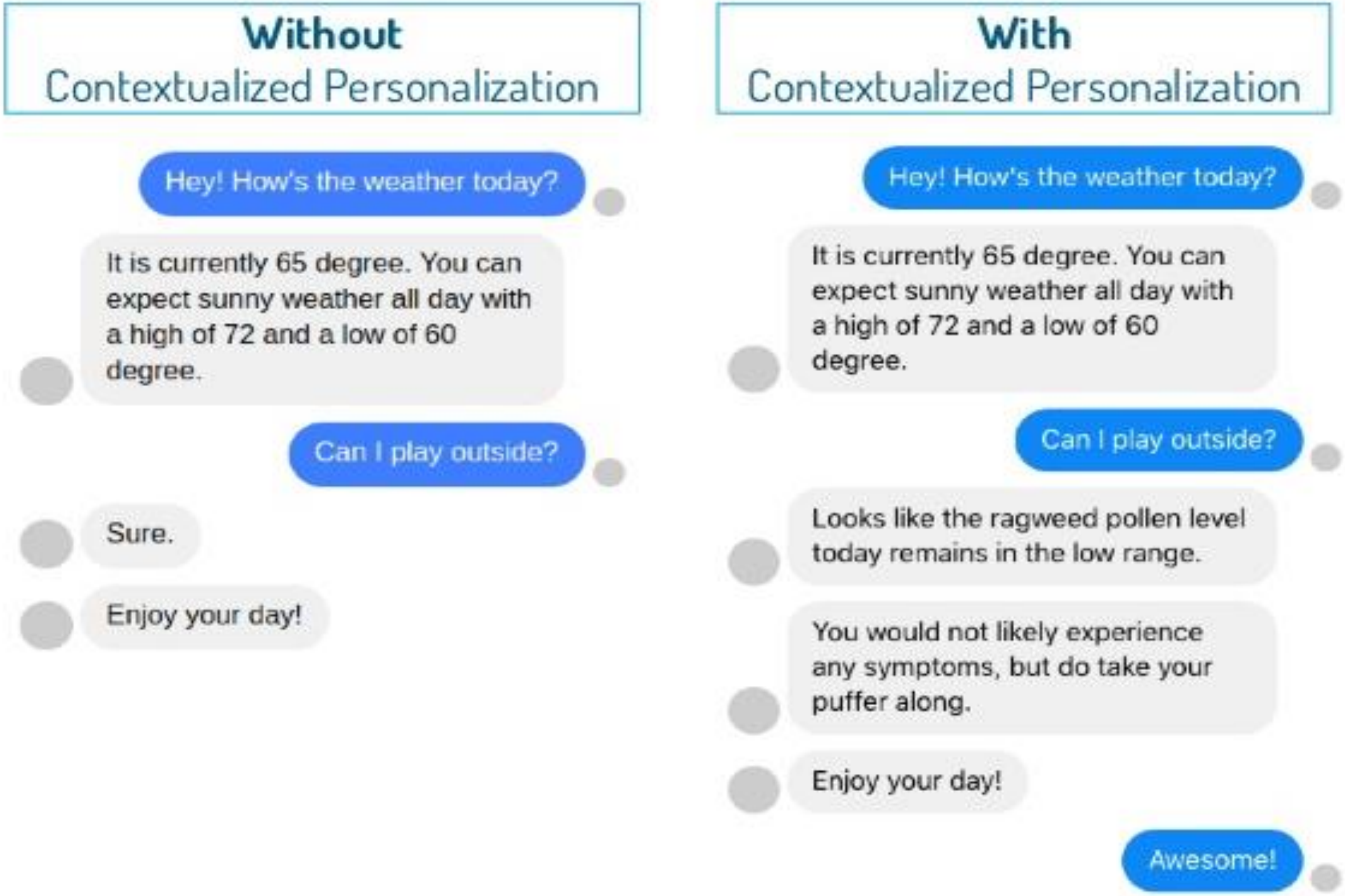
HOTEL Chatbot

Marriott International's

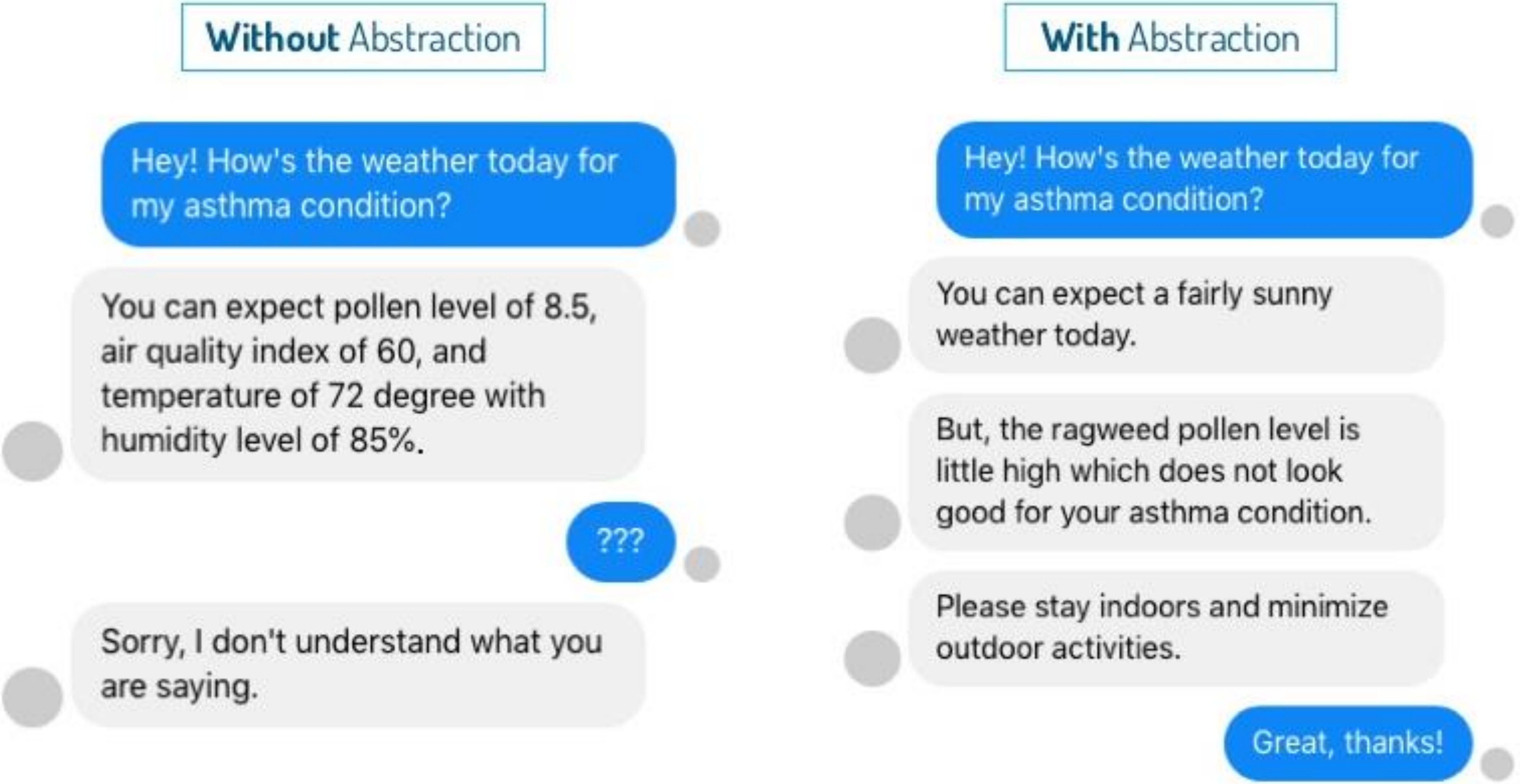
Book travel in more than 4,700 hotels

https://news.marriott.com/2017/09/marriott-internationals-ai-powered-chatbots-facebook-messenger-slack-alofts-chatbotlr-simplify-travel-guests-throughout-journey/


Without Domain Knowledge
I am your Infobot, ask me anything!
Tell me more about bupe.
I'm not sure I understand exactly what you're asking for.
Subutex?
Sorry, I didn't get you. Please try again.
I meant Buprenorphine!!
With Domain Knowledge
I am your Infobot, ask me anything!
Tell me more about bupe.
Bupe or Buprenorphine, sold under the brand name Subutex, is an opioid used to treat opioid addiction, acute pain, and chronic pain
Would you like to know more?
Sure!
Its side effects may include respiratory depression (decreased breathing), sleepiness, adrenal insufficiency, QT prolongation, low blood pressure, allergic reactions, and opioid addiction.



Source: https://www.slideshare.net/apsheth/humanlike-chatbots-promises-challenges-and-implications-102794389


Without Abstraction
Hey! How's the weather today for my asthma condition?
You can expect pollen level of 8.5, air quality index of 60, and temperature of 72 degree with humidity level of 85%.
???
Sorry, I don't understand what you are saying.
With Abstraction
Hey! How's the weather today for my asthma condition?
You can expect a fairly sunny weather today.
But, the ragweed pollen level is little high which does not look good for your asthma condition.
Please stay indoors and minimize outdoor activities.
Great, thanks!

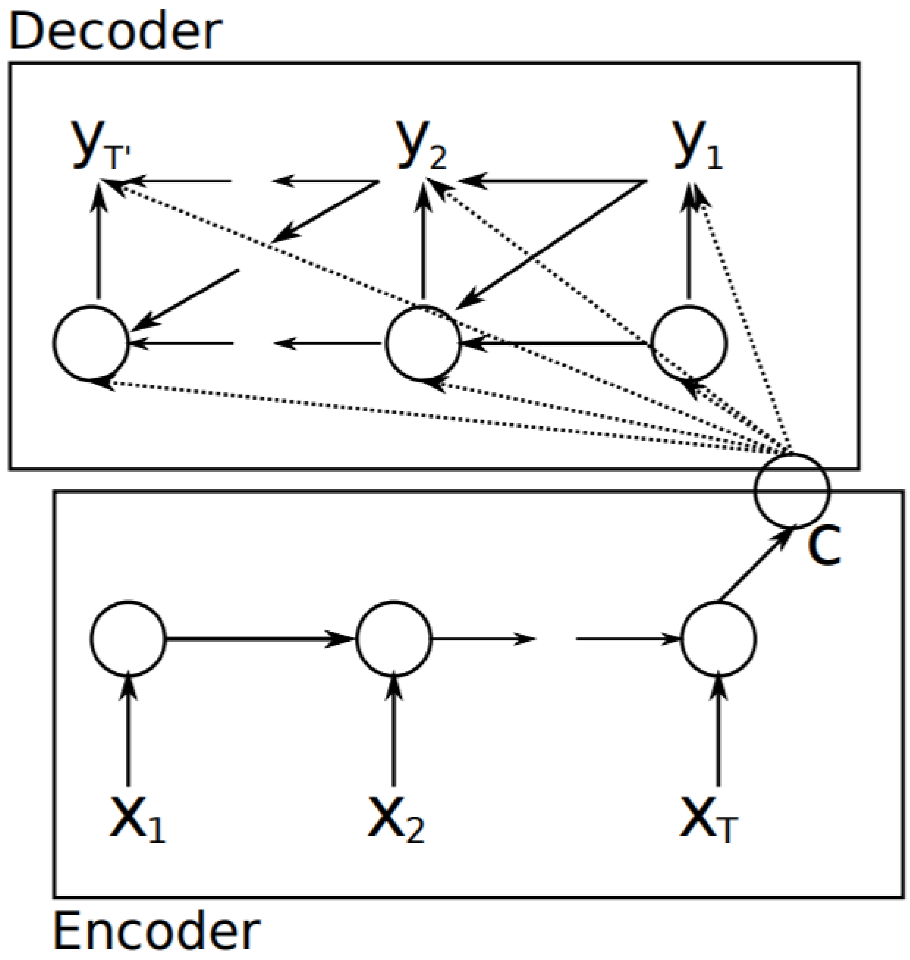
Literature Survey

Source: Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation **[s2s]**

Kyunghyun et al. 2014

Theano (official), Pytorch, Tensorflow

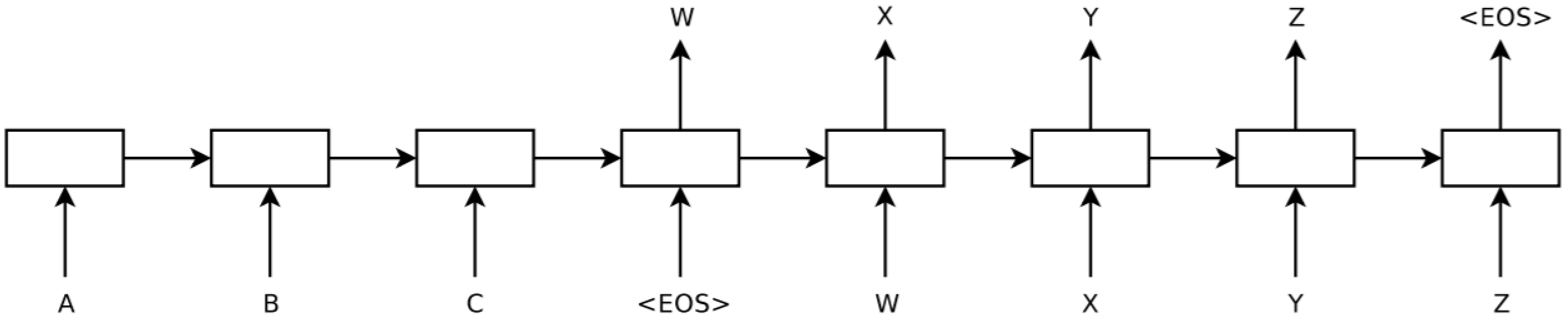
- Two RNNs for encoding and decoding of sequences, jointly trained
- Equations in the paper using gated recurrent unit
- They only looked at rescoring translation phrases, not generating



Literature Survey

Sequence to Sequence Learning with Neural Networks [s2s]
Sutskever et al. 2014
Keras, Numpy, Tensorflow

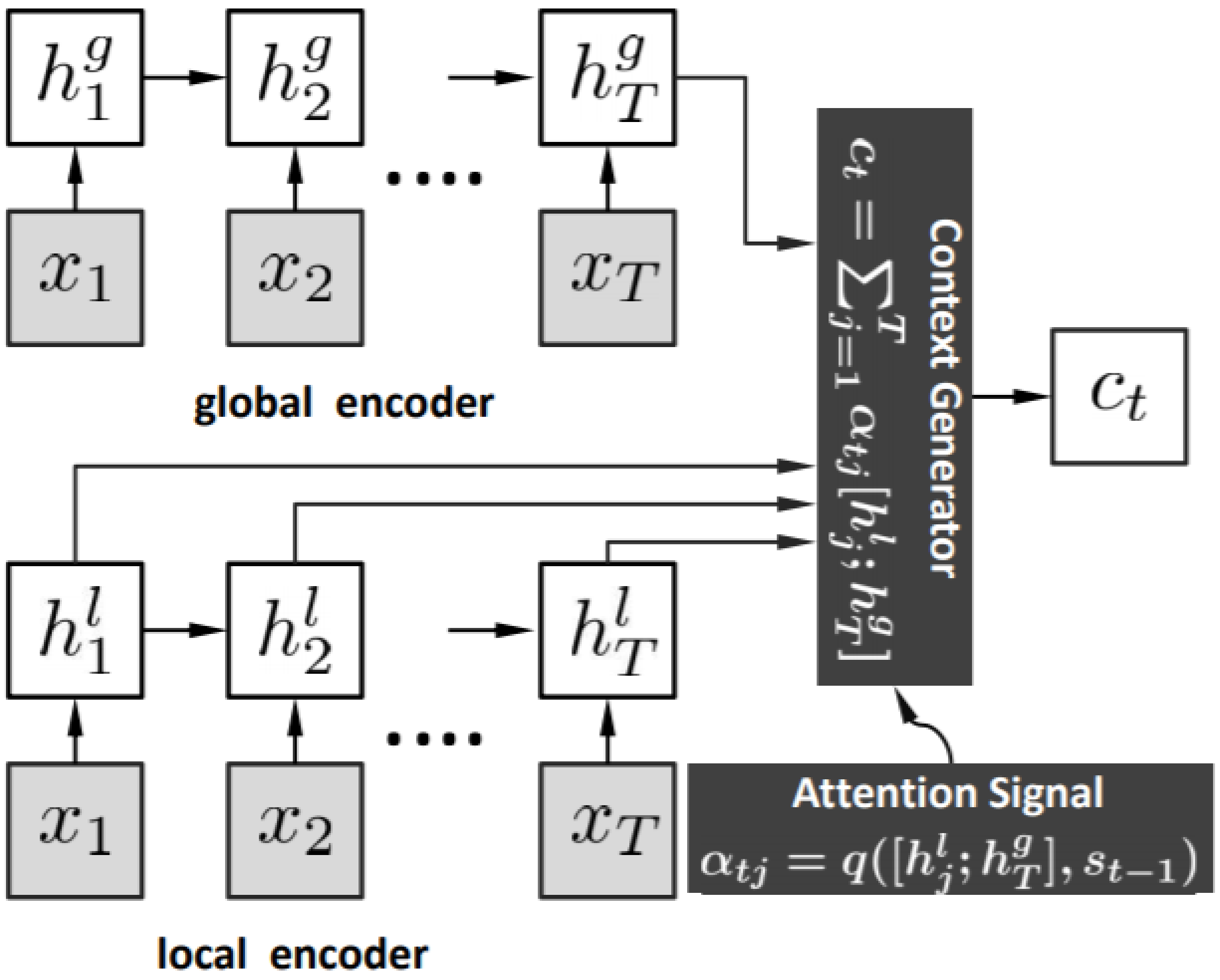
- Encoder-decoder with LSTM (pretty big architecture)
- Words are reversed in source sequence for better performance
- Left to right beam search decoder



Literature Survey

Neural Responding Machine for Short-Text Conversation [chat]
Shang et al. 2015

- Encoder-decoder model applied to twitter style 2-turn conversations, with bahdanau attention and GRU and beam search for decoding
- Combines the bahdanau attention model with the original global context vector representation
- Evaluation done with human judgment



Literature Survey

A Neural Conversational Model [chat]
Vinyals et al. 2015
Torch, Keras, Theano, Tensorflow

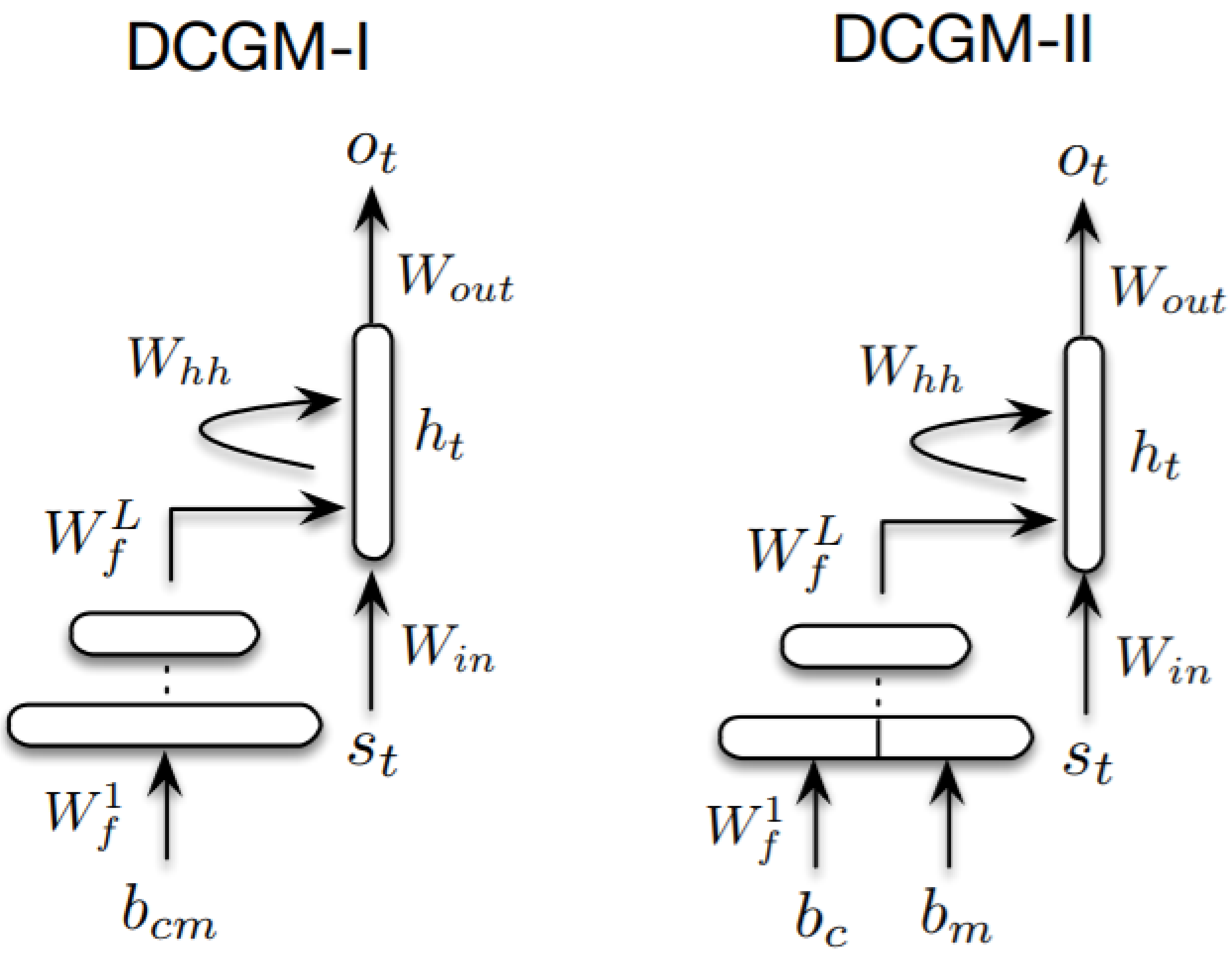
- IT helpdesk dataset and movie subtitles; Big architectures and big vocabs
- Input sequence is what has been conversed so far (context), output sequence is the reply
- Objective function optimized is not the actual objective achieved through human communication
- Problem mentioned is with the inconsistent answers (there is no personality) and with not being able to evaluate correctly

Literature Survey

[A Neural Network Approach to Context-Sensitive Generation of Conversational Responses]() [chat]

Sordoni et al. 2015

- Encode past information, which is then decoded to promote responses
- Separate context from last message
- They use IR to generate more responses to a (c,m,r) triple based on bag of words
- They use a ton of features together with the neural network models to generate likely responses



Literature Survey

How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation [chat]

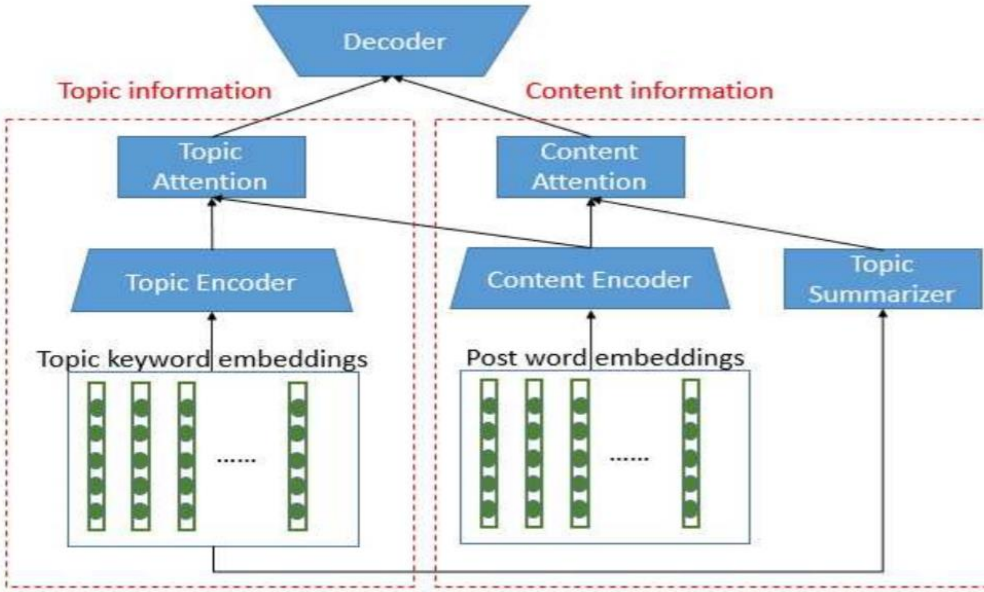
Liu et al. 2016

- BLEU not good where responses are diverse with no matching words, deltaBLEU is weak and needs human annotation for multiple reference replies
- BLEU is based on n-grams, METEOR produces alignment between response and ground truth, ROUGE is based on longest common subsequence
- Greedy matching is based on matching words with closest embedding vectors in response and truth, embedding average: sentence level embedding
- They all correlate (with human judgment) poorly on twitter dataset and not at all on ubuntu dataset

Literature Survey

Topic Aware Neural Response Generation [chat]
Xing et al. 2016
Theano

- Represent people's prior knowledge about the topic, and embed this into reply of seq2seq model with attention
- Two encoders with separate attention modules, one is bidirectional RNN, other is for topic words, then their attention is jointly fed into decoder
- The two encoders can affect each others attention, topic attn finds relevant info, content attn determines the content focus
- Topic word list obtained from twitter LDA model, they play the role of classification and association in response generation (better first words chosen)

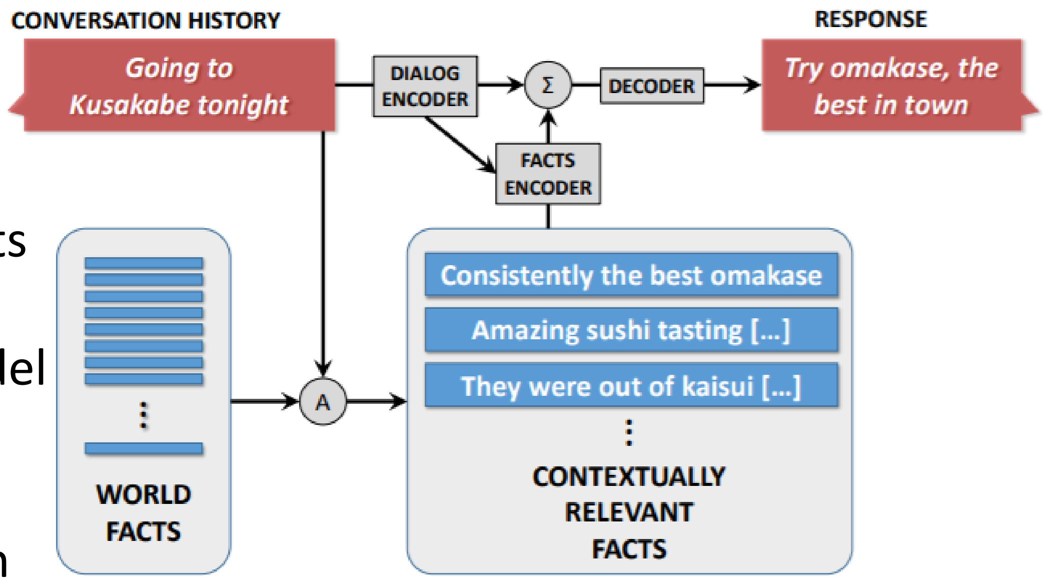


Literature Survey

A Knowledge-Grounded Neural Conversation Model [chat]

Ghazvininejad et al. 2017

- Condition responses based on conversation history and external facts (amazon, wikipedia) relevant to current context
- NER is used for example to make a query to retrieve facts; these are fed into a fact encoder -> this summed with conversation encoder are fed into decoder
- Fact encoder is similar to memory network, retrieves and weights facts based on user input and conversation history
- Multitask learning: first task is conversational, pure enc-dec model trained; second task exposes the full model to facts as well; third task is similar to autoencoder, it uses facts for both encoders
- Twitter dataset with mentions of local business, augmented with facts (foursquare tips): many contextually relevant facts -> filter them with tf-idf, retain 10 tips
- They use beam search and N-best lists reranking based on MMI
- The results are somewhat more diverse than baseline seq2seq



Conversation **Intent**

Natural Language Understanding (NLU) Techniques

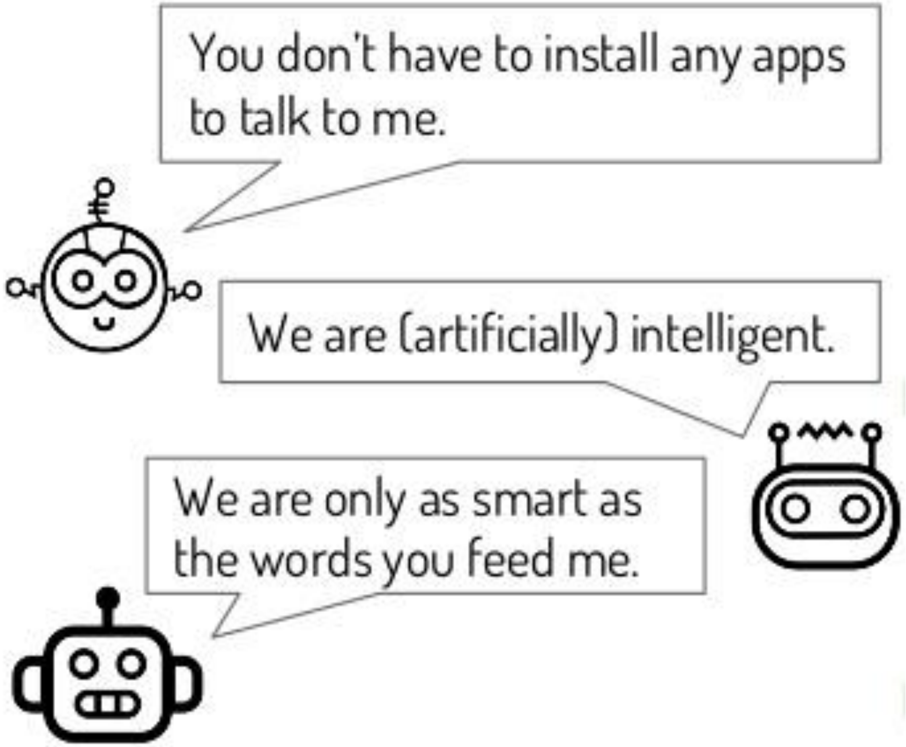
- Named-entity Recognition (NER) and Disambiguation
- Sentence Completion
- Topic and Domain Detection
- Implicit Entity Recognition
- Relation Extraction
- Text Summarization
- Sentiment, Emotion, and Intent Detection
- Emoji Sense Disambiguation
- Machine Translation
- Ranking and Selection (Open-domain social conversations)

Inappropriate and Offensive Speech Detection

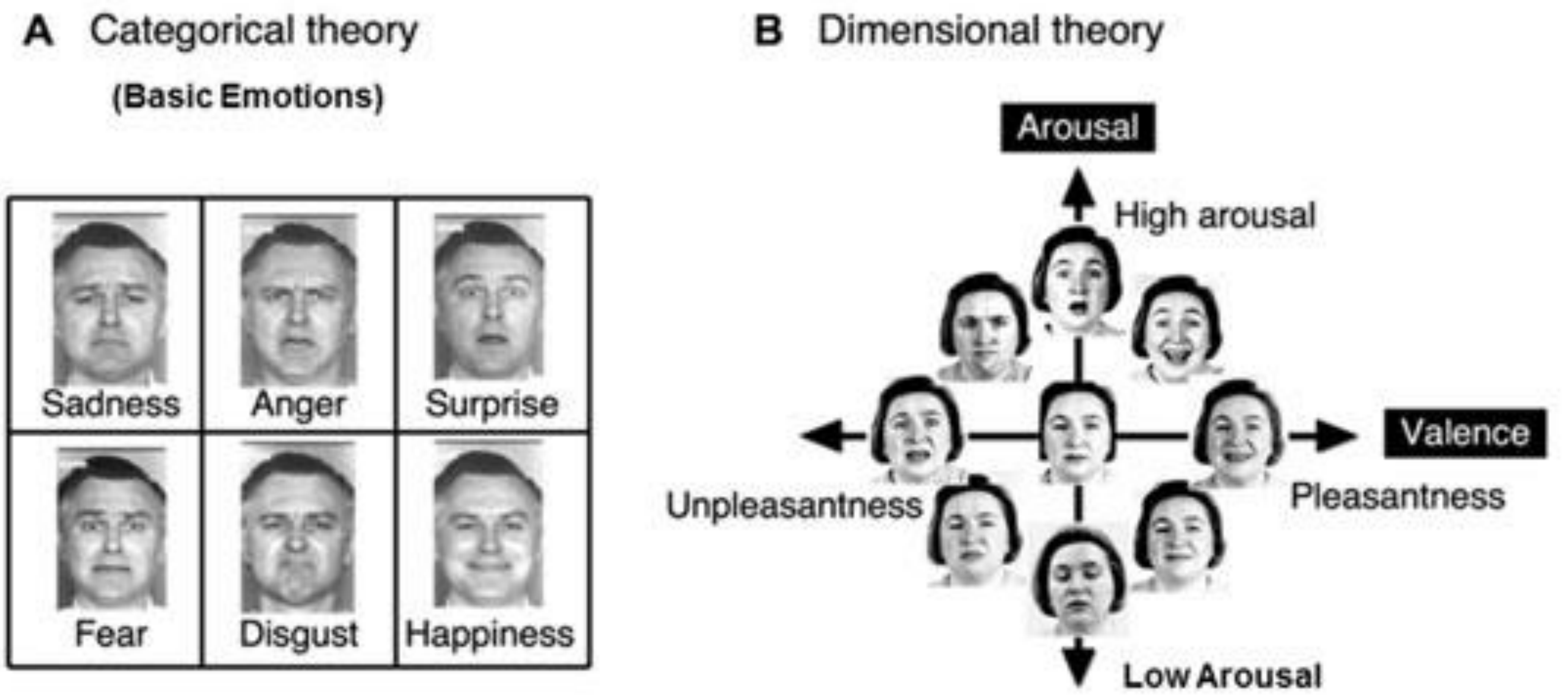
Conversational Datasets, Commonsense Reasoning and Knowledge Ingestion

Conversational Topic Tracker

Response Generation



Emotion Theory



Reference: Matsuda et al. "The implicit processing of categorical and dimensional strategies: an fMRI study of facial emotion perception." *Frontiers in human neuroscience* 7 (2013).

Do you like Pineapple on Pizza?

Baseline System Network Architecture And Features

Model Details:

- Hierarchical Recurrent Encoder-Decoder (HRED) architecture for handling deep dialog context
- Multilayer RNN with GRU cells. First layer of the utterance-level encoder is always bidirectional.
- Thought vector is fed into decoder on each decoding step.

Reference: Zhou et al. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. AAAI (2018)

Baseline System Network Architecture And Features

Model Details:

- Decoder can be conditioned on any string label. For example: emotion label or id of a person talking.
- The model is trained with **context size 3** where the encoded sequence contains **30 tokens or less** and the decoded sequence contains **32 tokens or less**.
- Both encoder and decoder contain **2 GRU layers** with **512 hidden units** each.
- Adding a second GRU layer to our network allows our model to capture higher-level interactions.

Reference: Sardoni et al. "A Hierarchical Recurrent Encoder-Decoder for Generative Context-Aware Query Suggestion.“ *CIKM 2015*

Baseline System Network Architecture And Features

Word Embedding Layer Details:

- Initialized using w2v model trained on the corpus.
- Embedding layer may either stay fixed or be fine-tuned along with all other weights of the network.
- Using these vectors is a form of *pre-training.* Intuitively, you are telling the network which words are similar so that it needs to learn less about the language.
- Using pre-trained vectors is particularly useful if you don’t have a lot of data because it allows the network to generalize to unseen words.

Baseline System Network Architecture And Features

Decoding Details

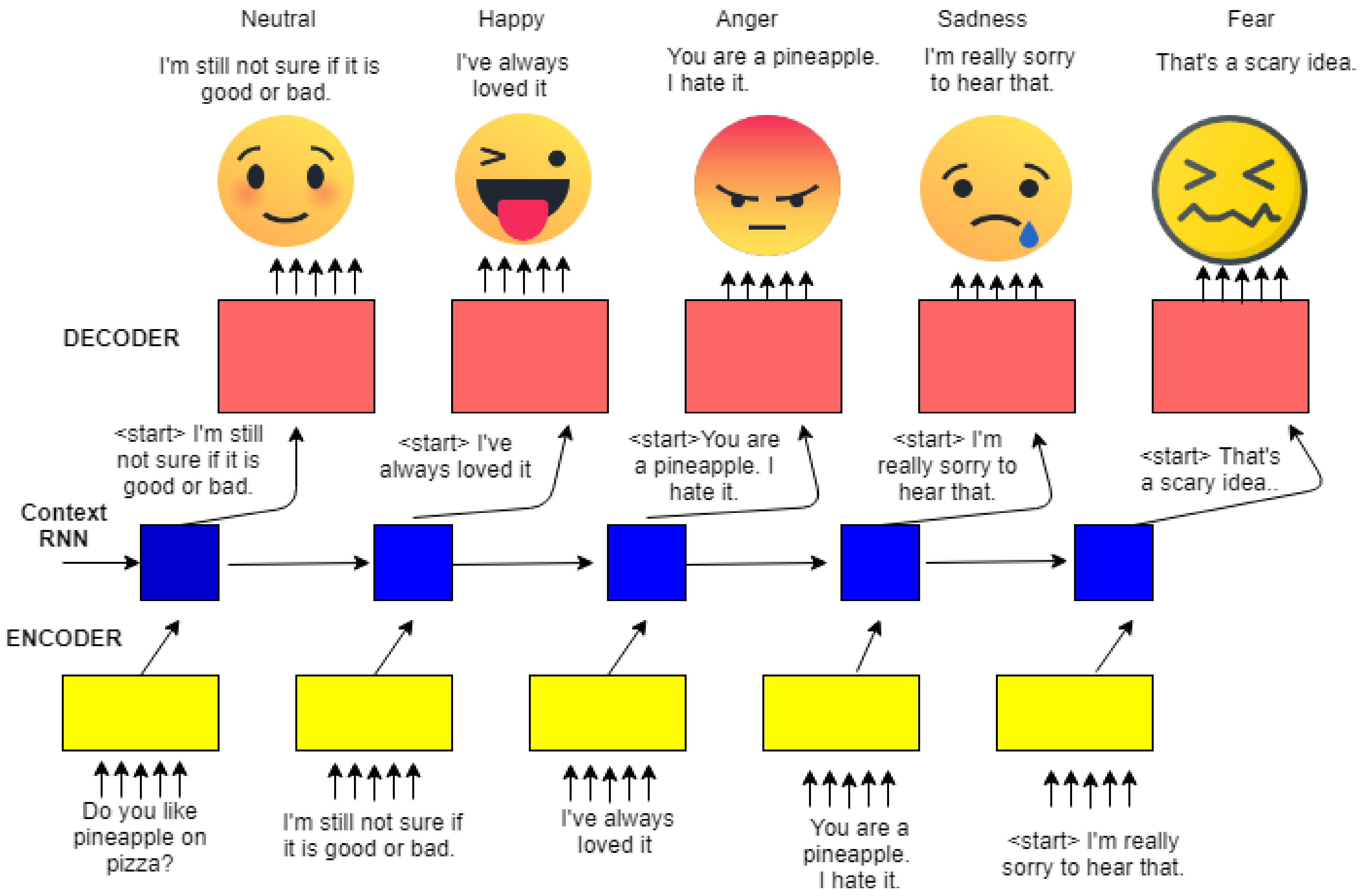
- 4 different response generation algorithms: "sampling", "beamsearch", "sampling-reranking" and "beamsearch-reranking".
- Reranking of the generated candidates is performed according to the log-likelihood or MMI-criteria

Metrics

- Perplexity
- n-gram distinct metrics adjusted to the samples size.
- Lexical similarity between samples of the model and some fixed dataset. Lexical similarity is a cosine distance between TF-IDF vector of responses generated by the model and tokens in the dataset.
- Ranking metrics: mean average precision and mean recall@k.
- Human Evaluation

Reference: Lie et al. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. *EMNLP 2016*: 2122-2132

Network Architecture And Features

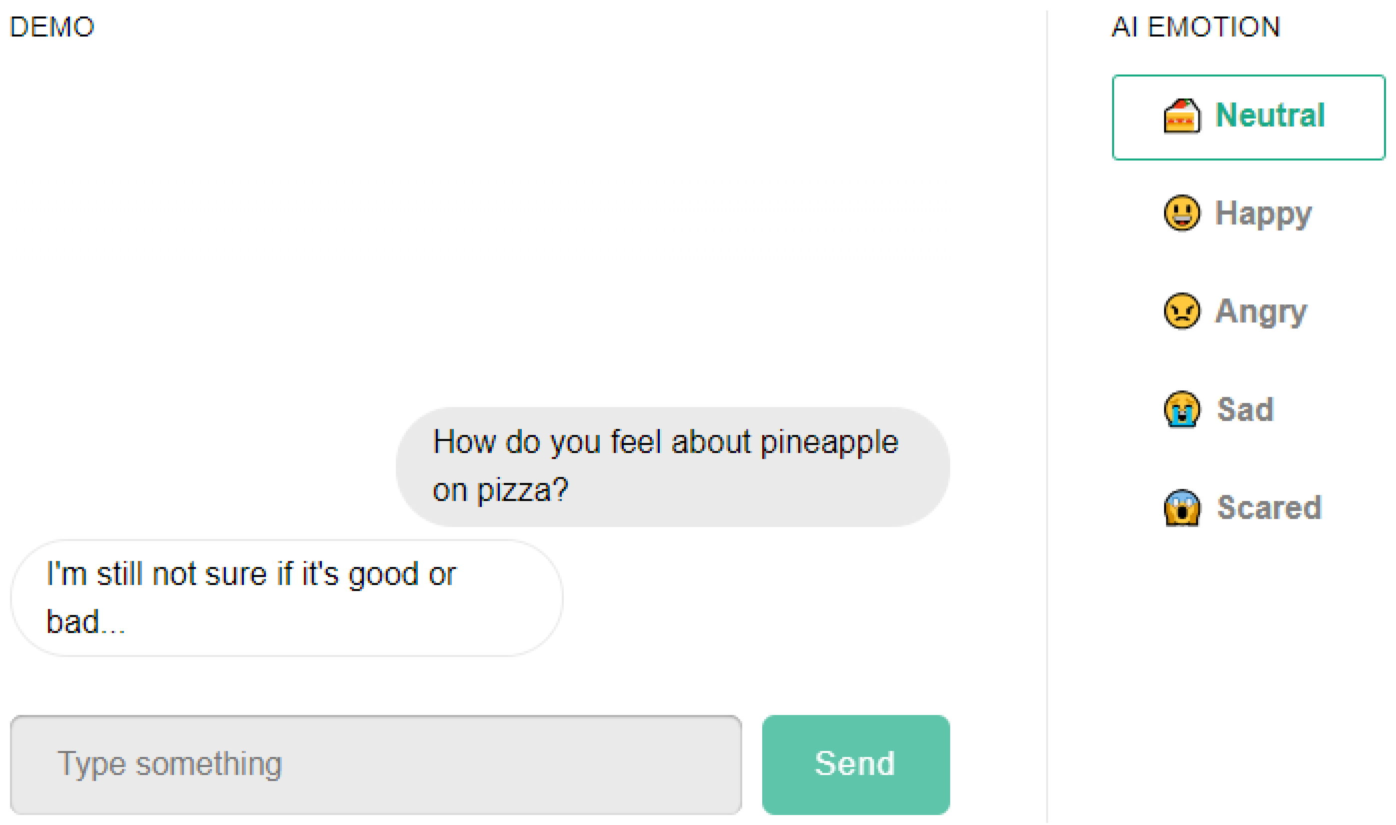


Network Architecture And Features

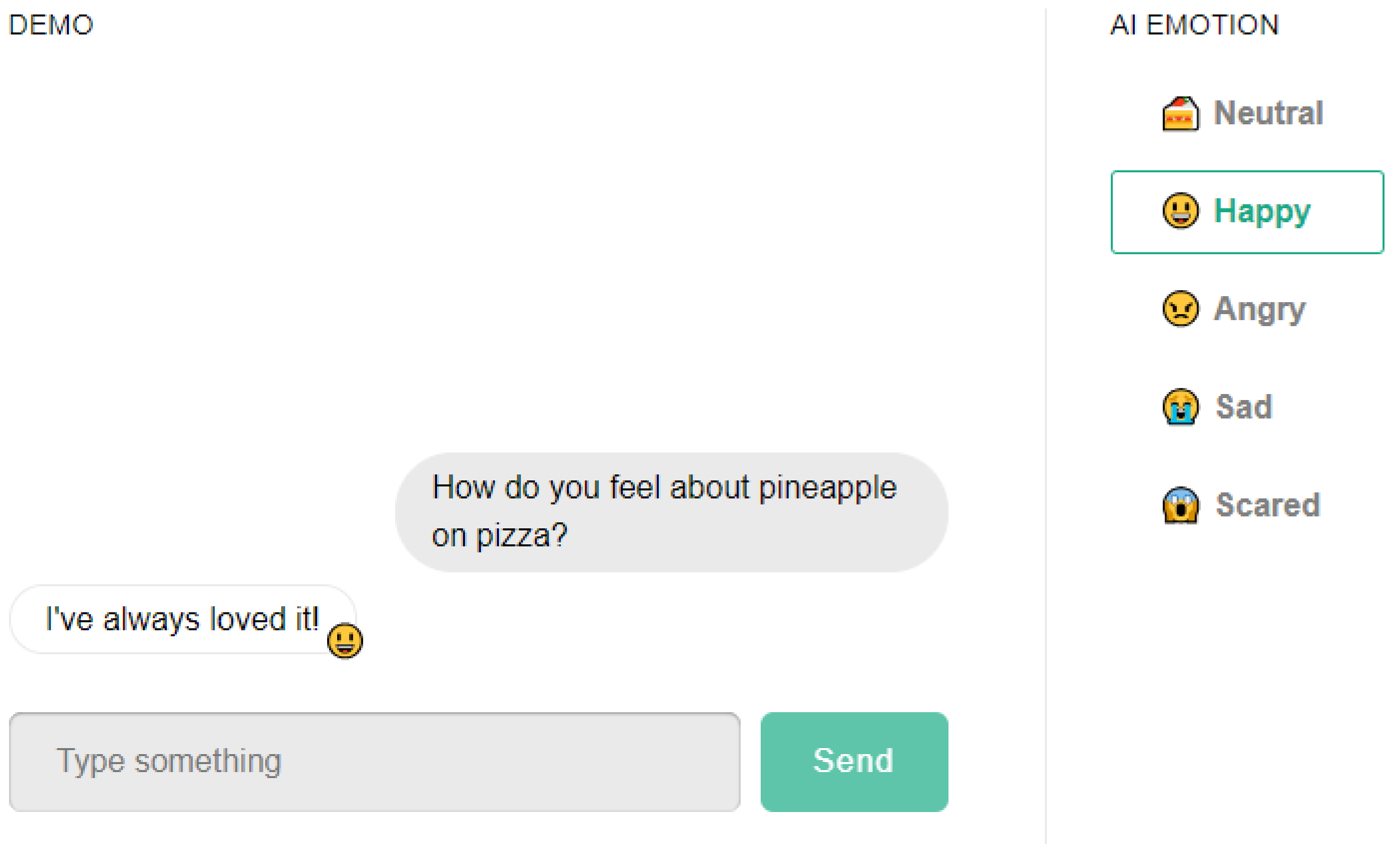
Data Set Details:

- Twitter preprocessed conversational data.
- To clean up the data, removed URLs, retweets and citations.
- Removed mentions and hashtags that are not preceded by normal words or punctuation marks and filtered out all messages that contains more than 30 tokens.
- Marked out each utterance with our emotions classifier that predicts one of the 5 emotions: "neutral", "joy", "anger", "sadness" and "fear".
- To mark-up our own corpus with emotions used DeepMoji tool

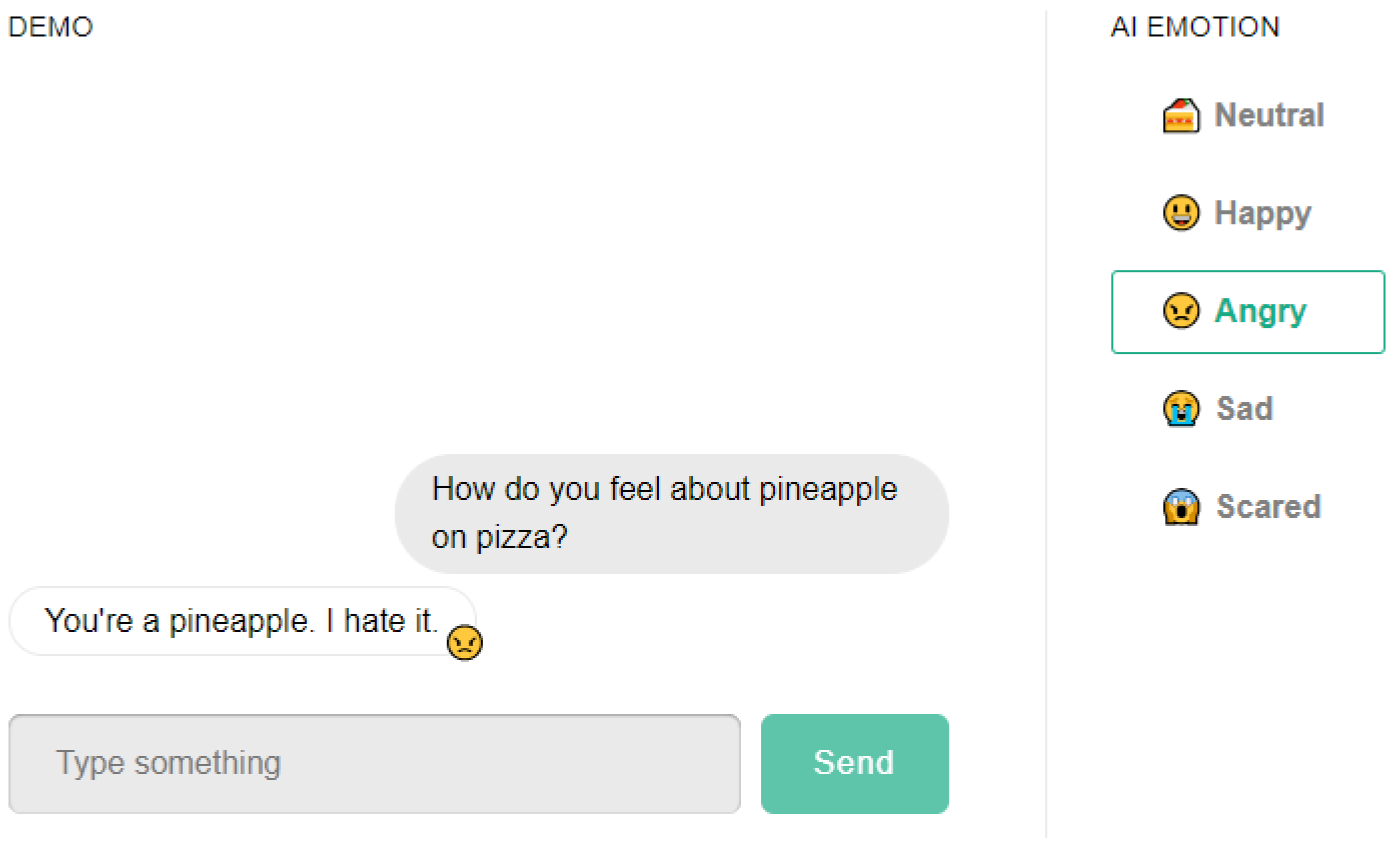
Network Architecture And Features



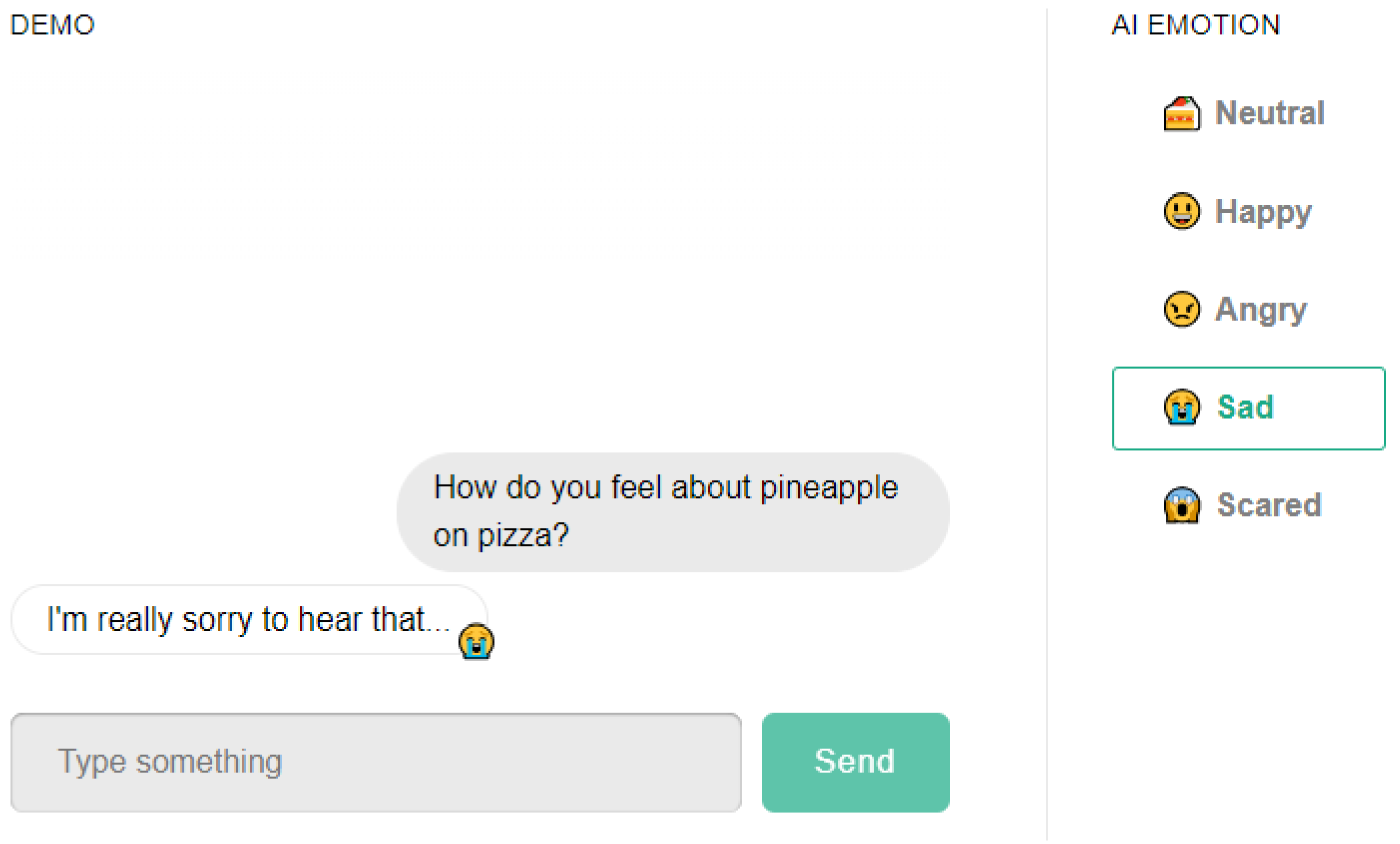
Network Architecture And Features



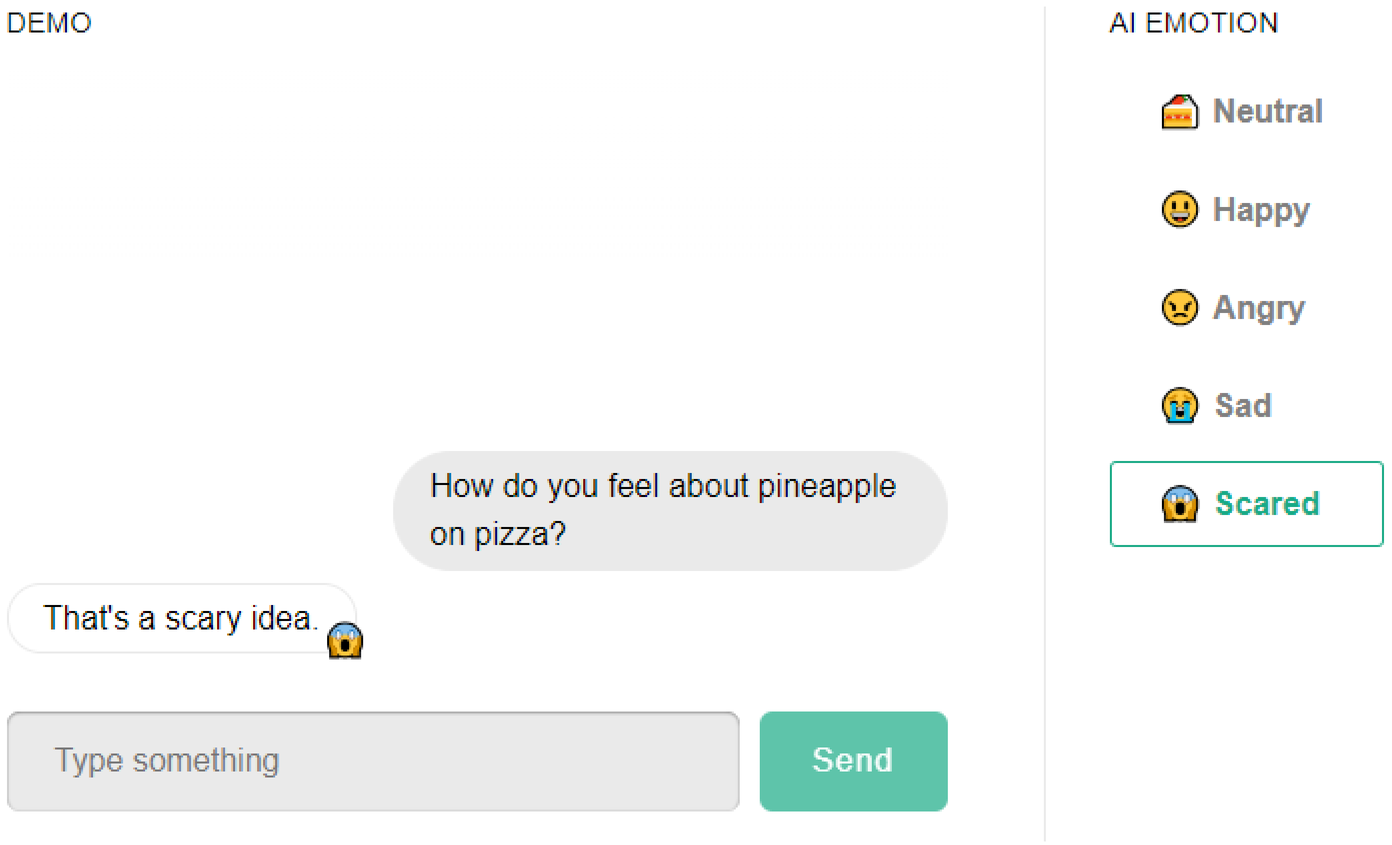
Network Architecture And Features



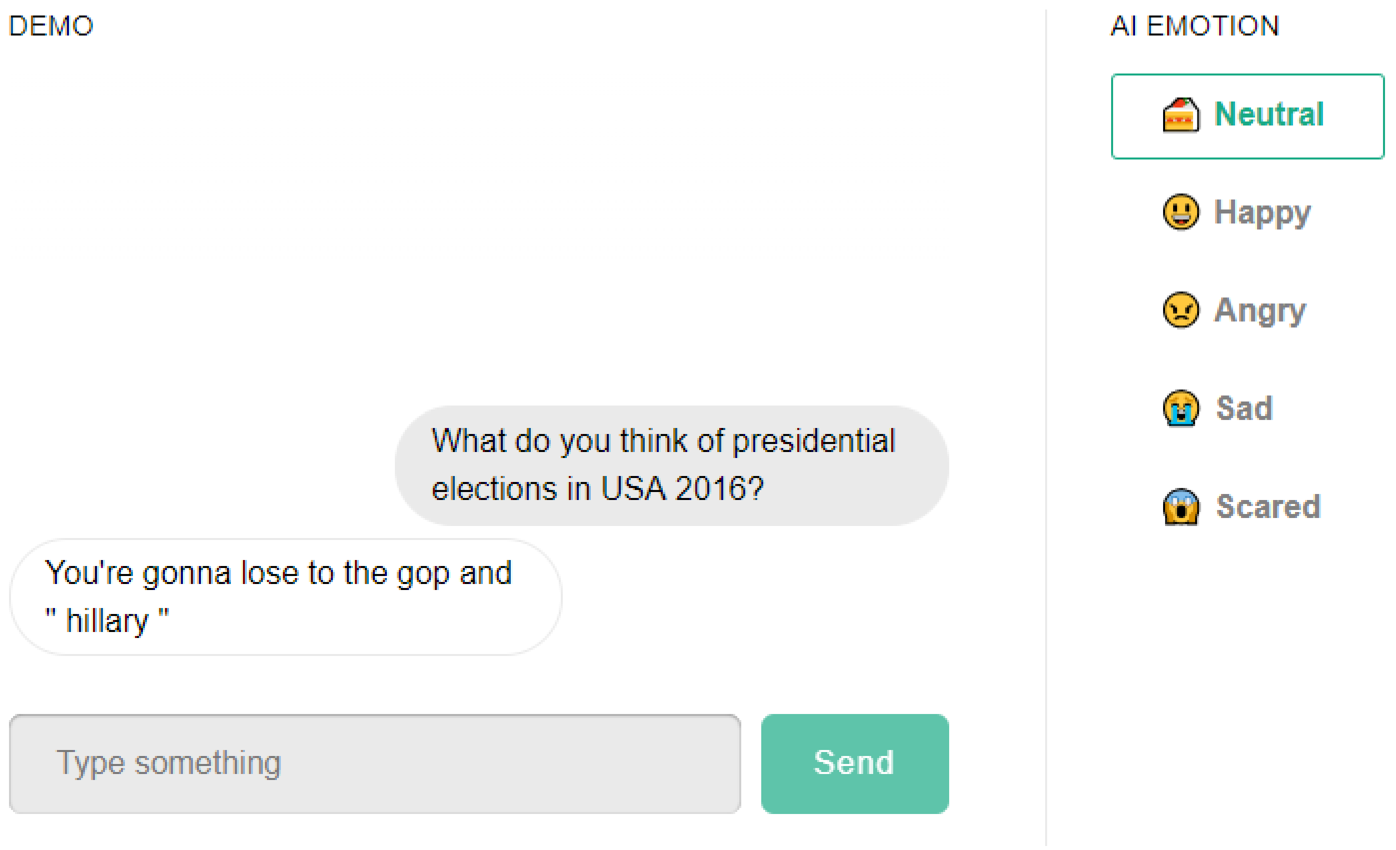
Network Architecture And Features



Network Architecture And Features



Examples From Demo (Contextual Understanding)



Baseline System Demo

Improving The Baseline System (My Next Steps)

- **Current Status:**

  Only Context and categorical emotion score are used as feature set.

- **Further Improvements:**
  - ➢ Improve the Feature Set: Semantic Role Modeling as another possible feature.
  - ➢ Assign Dimensional Emotion Score to the Training Corpus. User can choose two emotions at a time and also define the percentage of each emotion in the bot. Ex: 50% Happy and 50% Sad.
  - ➢ Build the new training corpus for the same.
  - ➢ Include Topic Based Response Generation i.e. incorporating emotions for domain specific chatbot response generation.
  - ➢ Incorporating mood of the user to decide independently the mood of the chatbot.
  - ➢ Incorporating persona in a chatbot. Ex: Your chatbot could be Yoda from Star Wars.

Reference: Li et al. Persona based neural conversational model ACL *2016*
Xing et al. Topic aware neural response generation AAAI 2017

Tentative Timeline

August 24, 2018	Master Seminar
September 30, 2018	Complete the implementation of the improvements
October 30	Collect Experimental Evaluation
November 30	Give Master Thesis Colloquium
December 30	Final Drafting and Submission of my thesis
January 15	Submit a paper (Hopefully!)
GRADUATE AND JOIN MY JOB AT MICROSOFT AS A DATA SCIENTIST!	

Summary

- Discussed the motivation for a chatbot and the Emotion Theory

- Discussed the architecture of the baseline system and played with its demo

- Discussed the next steps which briefly include:
  - ➢ Assign Dimensional Emotional Score To the Training Corpus to assign hybrid set of emotion range for chatbot responses
  - ➢ Giving Persona To The ChatBot
  - ➢ Taking Mood Of The User
  - ➢ Improving the Contextual Goals
  - ➢ Topic Aware Response Generation