Conversational Bots for Customer Support

Michal Shmueli-Scheuer, Ph.D.

IBM Research AI – Haifa

Motivation

80% of businesses want virtual agents by 2020

85% of customer service interactions will be powered by virtual agents in 2020

- ✓ Advances in AI and NLP
- ✓ Rise in messaging applications over social media

- ✓ Service that's always-on
- ✓ 60% – 80% cost saving over outsourced call centers powered by humans
- ✓ Personalization is key
- ✓ Scalability
- ✓ …

In this Talk..

- Affect in customer support conversations
  - Introduction
  - Human-2-bot

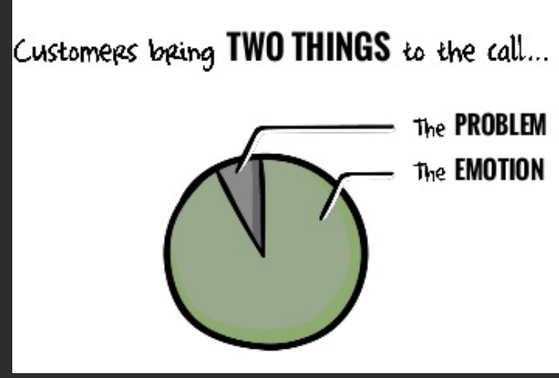
- Egregious conversations with bots

Affect

Openness
Consciousness
Extraversion
Agreeableness
Neuroticism

Happiness
Sadness
Anger
Confusion
Frustration
Disappointment
Gratitude
Hopefulness
Politeness

Affect in Customer Support



- Emotions are a cardinal aspect of customer service, as they relate directly to customer satisfaction and experience.
- How customers feel about an experience heavily influences their future purchasing behavior
  - Forrester has found that negative emotions have a much bigger impact than positive ones. Make a customer angry and they're not only likely to stop doing business with you, they'll tell everyone they can about their negative experience.
  - Positive experiences can go a long way when it comes to creating the coveted "promoter" customer.
  - 88% of customers will take positive action if they have a good experience with your business.

- Analyzes customer emotions during conversation and uses them to predict and optimize the interaction.

Affect and Conversational bots

- We believe that virtual agent should have a cognitive layer of affect while interacting with customer.
- Previous researches showed that even when customer knows that they are interacting with virtual agent, they still expect this agent to behave like human.
- Thus we want automated conversation agents to be able to detect and express affect.
- Affect can be used in the agent assignment process (to assign an agent with personality traits that would maximize customer satisfaction).
- Mattersight.com

- Nass, C., Moon, Y.: Machines and Mindlessness: Social Responses to Computers. Journal of Social Issues 56(1), 81-103 (2000)
- Nicole C. Krämer, Social Effects of Virtual Assistants. A Review of Empirical Results with Regard to Communication, Proceedings of the 8th international conference on Intelligent Virtual Agents, September 01-03, 2008, Tokyo, Japan

Best Short Paper Award

Neural Response Generation for Customer Service Based on Personality Traits

iNLG 2017

Joint work with Jonathan Herzig, Tommy Sandbank, David Konopnicki

Motivation

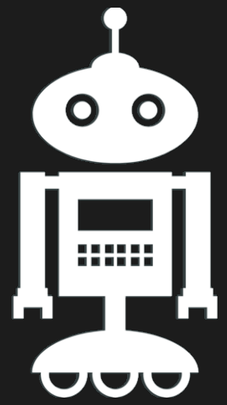
- Automated conversational agents are becoming popular for various tasks

Customer Service

Mobile Apps

Messaging Channels

Internet-of-Things



Robots

Motivation

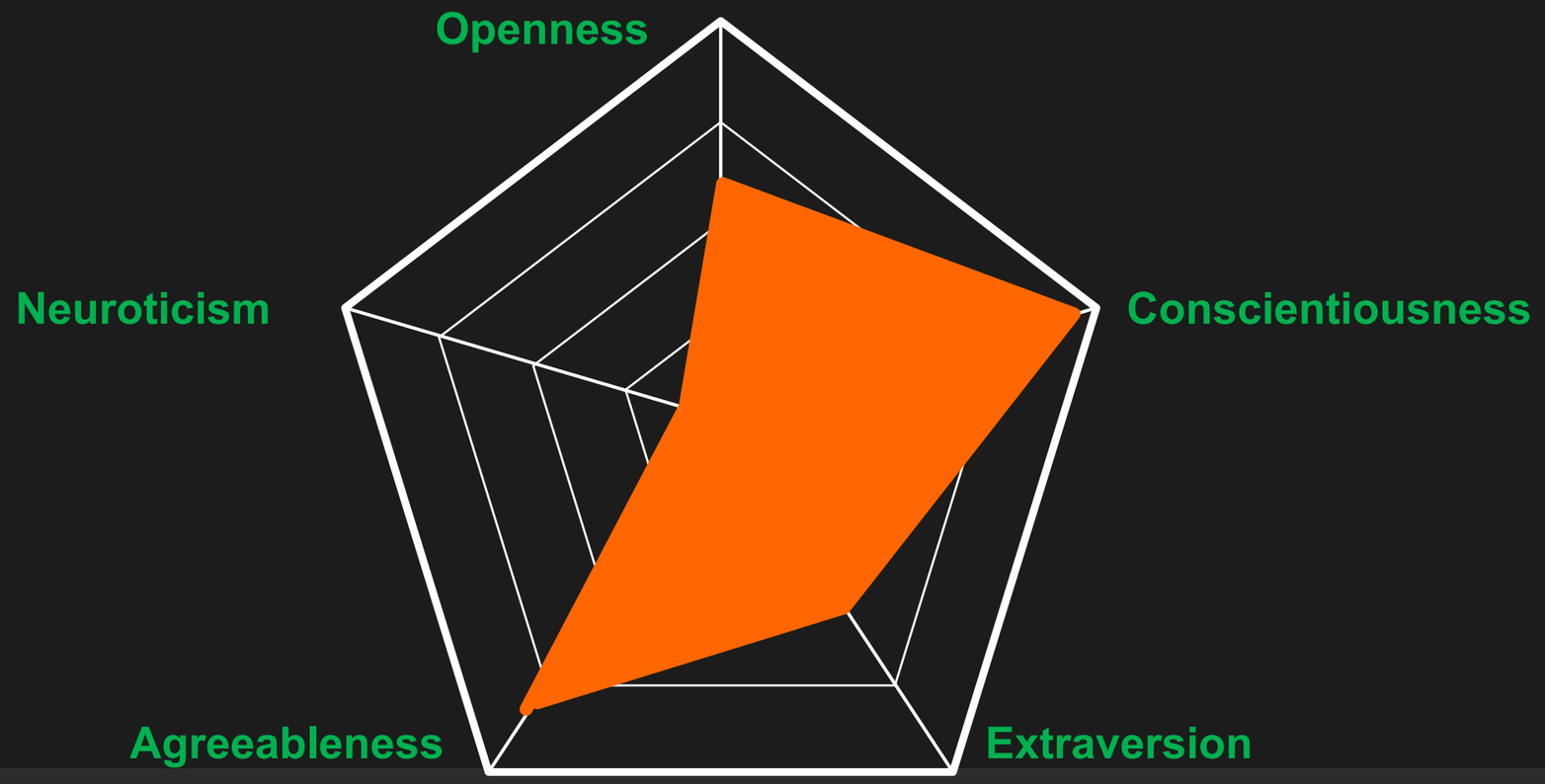
- Automated agents benefit from adapting their personality according to the task at hand
  - Application
  - Brand
  - Culture

Personality Traits

- A set of traits which represent durable characteristics of a person
- The Big Five model:
  - Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism
- Agreeableness: a person's tendency to be compassionate and cooperative toward others
- Conscientiousness: a person's tendency to act in an organized or thoughtful way

Problem Description

- We study how to encode personality traits as part of neural response generation
- Our approach builds upon a sequence-to-sequence (Seq2Seq) architecture



Problem Description

Customer: Why isn’t your service working???

Consensus-agent: We are aware of the issue and are working to fix it.

Agreeableness-agent: We’re here to help! Are you getting any error messages or codes?

Customer: You guys messed up my delivery today.

Consensus-agent: Please contact us if you don’t get it by the end of the week.

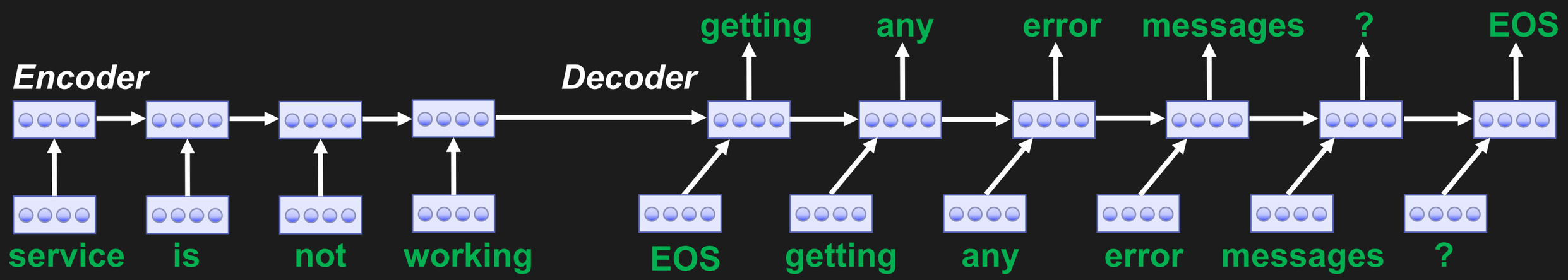
Conscientiousness-agent: Please email us with your tracking #, details and contact #. We’ll check on it.

Related Work

- PERSONAGE system (Mairesse and Walker, 2007):
  - Generation parameters: verbosity, exclamation, word length etc.
  - Rules to generate the desired outputs
  - Learn parameter values that correlate with personality trait intensity
- A persona-based neural conversation model (Li et al., 2016)
  - Learn persona vectors for each agent in the training
  - Abstract vectors, only for agents present in training data

Sequence-to-Sequence

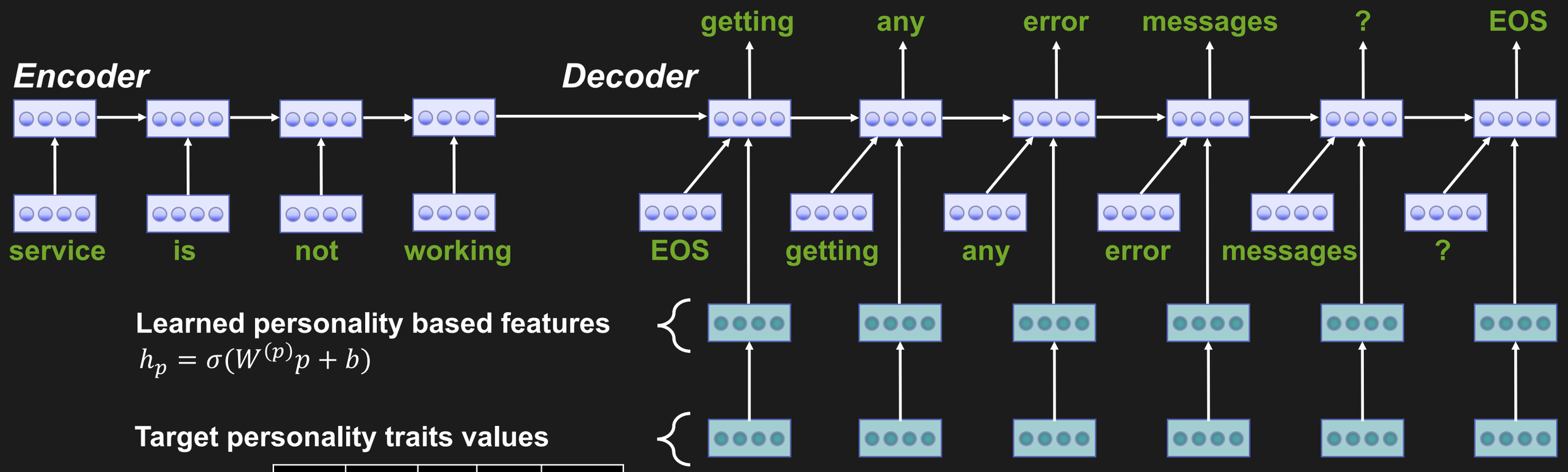
- LSTM encoder
- Attention based LSTM decoder



$$s_{j+1} = LSTM([\phi^{(out)}(y_j), c_j], s_j)$$

Our Personality-based Model

Hidden state update $s_{j+1} = LSTM([\phi^{(out)}(y_j), c_j, h_p], s_j)$



e.g., p=

0.85	0.85	0.3	0.5	0.15

Dataset

- We utilized the dataset presented in (Xu et al., 2017)
- 1M conversations over 62 customer service Twitter channels
- Filtered customer-agent utterance pairs for agents with less than 30 pairs
- 87.5K utterance pairs after filtering

Personality Traits Detection

- To extract personality traits for agents in our experiments we utilized **IBM Personality Insights**
- Given agent's text, infers three personality traits models

Model	Size	Traits
Big five	35	Agreeableness, Conscientiousness, Extraversion, Neuroticism, Openness
Needs	12	Excitement, Harmony, Curiosity, Ideal, Closeness, Self-expression, Liberty, Love, Practicality, Stability, Challenge, Structure
Values	5	Self-transcendence, Conservation, Hedonism, Self-enhancement, Excitement

Experiments and Results

- Warm start: Each agent has some data in train/dev/test

Model	Perplexity	BLEU
Seq2Seq	11.49	6.3%
PERSONA-BASED (Li, 2017)	9.25	15.55%
PERSONALITY-BASED	9.62	12.46%

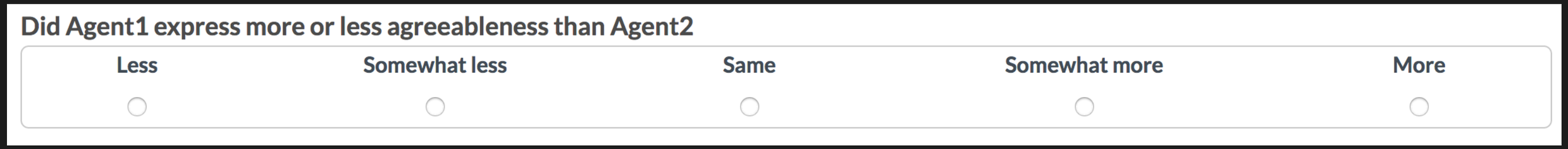
- Cold start: Data for some agent appears only in train or in dev/test

Model	Perplexity	BLEU
Seq2Seq	21.04	3.19%
PERSONALITY-BASED	19.64	4.67%

Human Evaluation

- Measure whether our model's output correlated with target personality traits
- Collected 60 customer utterances with negative sentiment

Customer	"App is broken and it's ruining my life"
Agent1	"I am sorry to hear this. Is there something that we can assist you with? we'd like to help in any way."
Agent2	"Sorry to hear this. Please get in touch with us through this link: www.company_a.com ."



- After averaging, 61% agreed with our model outputs (Somewhat or more)

Conclusion

- Target personality traits can serve as generation parameters to decoder
  - Experimented in customer service
  - Mild improvements in BLEU score
  - Validated by a human evaluation

Detecting Egregious Conversations between Customers and Virtual Agents

NAACL 2018

Joint work with Tommy Sandbank, Jonathan Herzig, David Konopnicki,
John Richards, David Piorkowski

e·gre·gious

/əˈgrējəs/

adjective

1. outstandingly bad; shocking.
   "egregious abuses of copyright"
   synonyms: shocking, appalling, terrible, awful, horrendous, frightful, atrocious, abominable, abhorrent, outrageous; More

Motivation

80% of businesses want virtual agents by 2020

85% of customer service interactions will be powered by virtual agents in 2020

Virtual agents may behave extremely bad which can lead to loss of customer loyalty and associated revenue…

Example

I got 2 quotes for the flight ticket, but I'm wondering what the details of each ticket are?

Please select ”Buy” next to the ticket you’d like to purchase.

No, I don’t want to buy yet till I know the details of the flights.

If you’re in the process of renting a car, please continue with by clicking ”Next”

Are you a real person ?

I am a digital assistant. I’ve been trained to answer questions about travels. Ask me any questions you have

I asked a specific question and you gave me a random answer about car rental which I’m not interested in

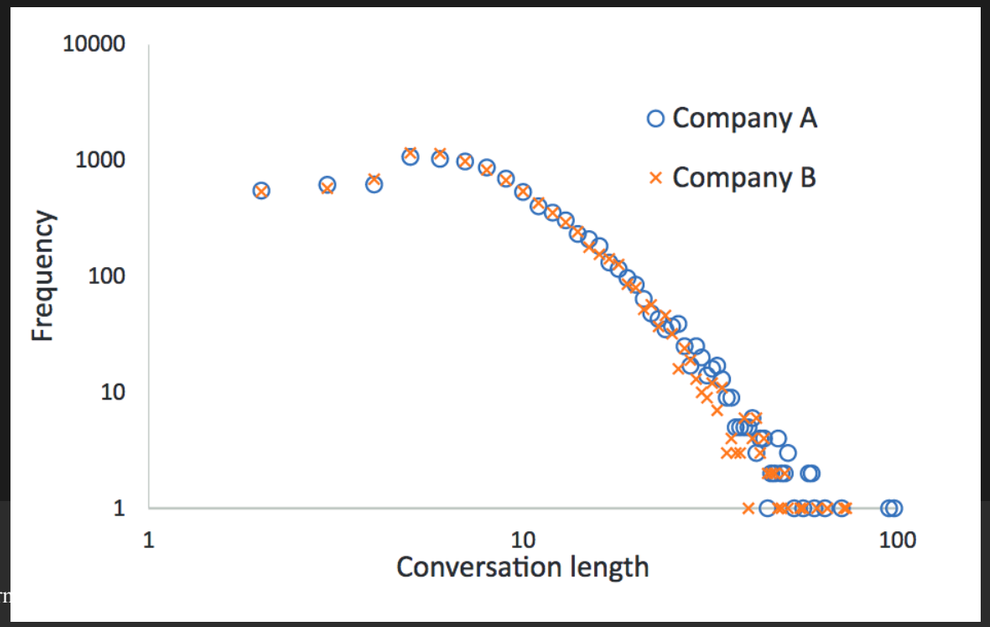
I’m not trained on that yet, but I’m still learning.
You may want to rephrase your question and try again

This service is pointless , can i talk to a real live person?

We don’t currently have live agents to chat with online

Dataset

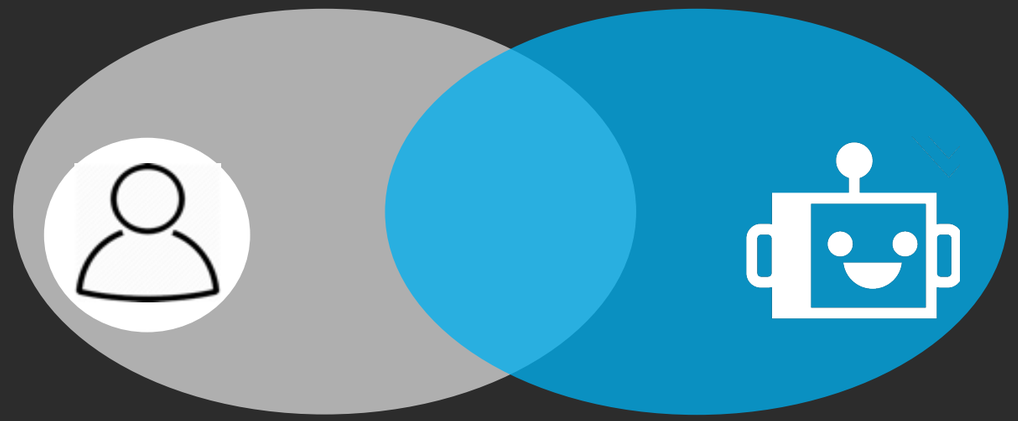
- We extracted data from two commercial systems that provide customer support via conversational bots
- Both agents are using similar underlying conversation engines, each has its own unique business logic
- From each system, we randomly extracted 10000 conversations



- On average 8.4 turns for Company A
- On average 4.4 turns for Company B

Methodology

- Treat the task as binary classification, where target classes are “egregious” or ”non-egregious”
- Classification is done at the end of each conversation (i.e., whole conversation)
- Extract 3 sets of features: agent response, customer input, customer-agent interaction
- Features can be contextual - where in the conversation they appear



Features – Agent response

Feature	Sub features	Contextual?	Description
Repeating Response Analysis	-	Yes	Similarity of subsequent agent responses
Unsupported Intent Analysis	-	No	Number of times the agent replied with “not trained” (or similar)

- Each sentence is represented by the averaging of its word embeddings (word2vec)
- Cosine similarity between sentences

Features – Customer input

Feature	Sub features	Contextual?	Description
Emotional Analysis	MAX NEG EMO	No	Max negative emotion in the conversation
	NEG SENT	No	Aggregated negative sentiment in the conversation
	DIFF NEG SENT	Yes/No	Difference between max turn-level negative sentiment and conversation-level
Rephrasing Analysis	#RPHRS	Yes	Number of customer rephrasing throughout the conversation
	RPHRS & NEG SENT	Yes	Rephrasing of subsequent turns with an average high negative sentiment
Asking for a Human Agent	HMN AGT & NEG SENT	No	Negative sentiment when asking for a human agent
Unigram Input		Yes	Turns that contained only one word

- NEG EMO - frustration, sadness, anger, confusion, etc.
- NEG SENT - summation of all NEG EMO

Features – Customer-Agent interaction

Feature	Sub features	Contextual?	Description
*** & Unsupported Intent Analysis**	NEG SENT & AGNT !TRND	No	Customer negative sentiment with agent replying “not trained”
	HMN AGT & AGNT !TRND	No	Customer asking to talk to a human agent followed by the agent replying “not trained”
	LNG SNTNS & AGNT !TRND	No	Customer long turn followed by an agent “not trained” response
Rephrasing Analysis & *	RPHRS & SMLR	No	The similarity between the customer’s turn and the agent’s response in case of customer rephrasing
	RPHRS & AGNT !TRND	No	The similarity between the customer’s turns when the agent’s response is “not trained”
Conversation length		No	Total number of customer turns and agent responses

Experimental Setting

- Sampled 1100 random conversations from Company A, and 200 random conversations from Company B.

- Each conversation was tagged by 4 HCI expert judges using this guideline

  "Conversations which are extraordinarily bad in some way, those conversations where you'd like to see a human jump in and save the conversation"

- Inter-rater reliability between all judges, measured by Cohen's Kappa = 0.72

Experimental Setting – cont.

- Company A - 95 were marked as egregious (8.6%)
- Company B - 16 were marked as egregious (8%)

- Baseline models
  - Text-based - unigrams, bigrams, NRC lexicon features, presence of exclamation marks, question marks, links, emoticons
  - Rule-based - virtual agent responded with a “not trained” reply, or occurrences of the customer requesting to talk to a human agent.

- Classifier was implemented using SVM with linear kernel

Results- Classification performance

	Egregious			Non-Egregious		
Model	P	R	F	P	R	F
Rule-based	0.28	0.54	0.37	0.95	0.87	0.91
Text-based	0.46	0.56	0.5	0.96	0.94	0.95
EGR	**0.47**	**0.79**	**0.59**	0.98	0.92	0.95

- Feature set contribution analysis

Results- Cross domain

- Models trained on company A's data, tested on company B's data

	Egregious			**Non-Egregious**		
Model	P	R	F	P	R	F
Rule-based	0.15	0.12	0.14	0.93	0.94	0.93
Text-based	0.33	0.06	0.11	0.92	0.99	0.96
EGR	**0.41**	**0.81**	**0.54**	0.98	0.9	0.94

Results- Customer Rephrasing Analysis*

- Natural language understanding (NLU) error - agent's intent detection is wrong, and thus the agent's response is semantically far from the customer's turn
- Language generation (LG) limitation – intent is detected correctly, but the customer is not satisfied by the response
- Unsupported intent error - customer's intent is not supported by the agent

	Egregious	**Non-Egregious**
NLU error	48%	48%
LG limitation	33%	37%
Unsupported intent error	18%	14%

*Sarikaya 2017, Sano et al. 2017

Conclusions & Future work

- Possible to detect egregious conversations using different features from the conversation
- Features are robust and can be applied in different domain

- Real time detection of egregious conversation
- Collect more data and using neural approaches
- Log analysis tool to explain the root causes

Thanks!
Questions?

For more information:

Michal Shmueli-Scheuer
shmueli@il.ibm.com