Current Topics in Data Engineering 2021
Evaluating chat bots (conversational agents)

Mohammad Tanzil Alam
Instructor Name: Prof. Dr. Adalbert F.X. Wilhelm

JACOBS UNIVERSITY

Abstract: Open domain dialog system are difficult to manually evaluate hence different approaches were used to evaluate the response. ChatEval aims scientific framework similar to Appraise. Different approaches were used to evaluate the response like Cohen's kappa and Item Response Theory (IRT) being the best approach in which we will be looking further.

Introduction/Motivation: Appraise is an open source tool which use to manually evaluate **Machine Translation** output.[1]

Figure 1: Human Evaluation [2]

DBDC(Dialogue breakdown detection): Detect whether the conversation between human and system is getting a dialogue breakdown.

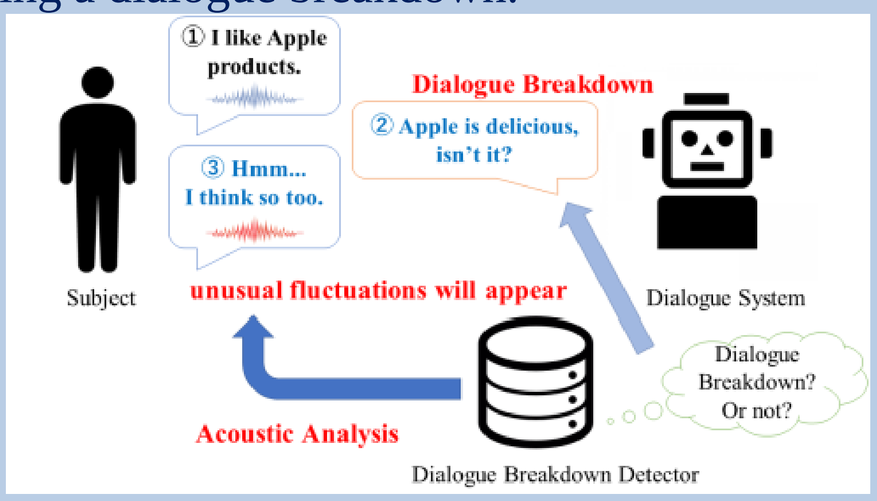


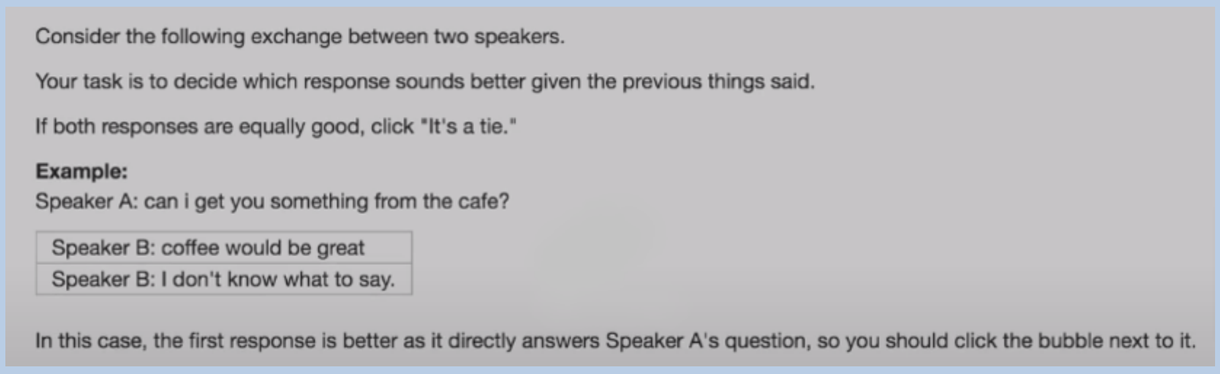
Figure 2: Example of DBDC [3]

Method:

Cohen's kappa: 0.21-0.60 fair to moderate : It is used to measure interrater reliability (to which extent data collected in study are correct w.r.t variable measured) to find the IAA a.k.a Inter Annotation Agreement[4]

Item Responsive Theory(IRT): IRT is use for evaluating difference between different models and quality of the evaluation dataset [5]. IRT results in highly interpretable estimations with less noisy judge than previously proposed method. [6]

Example:



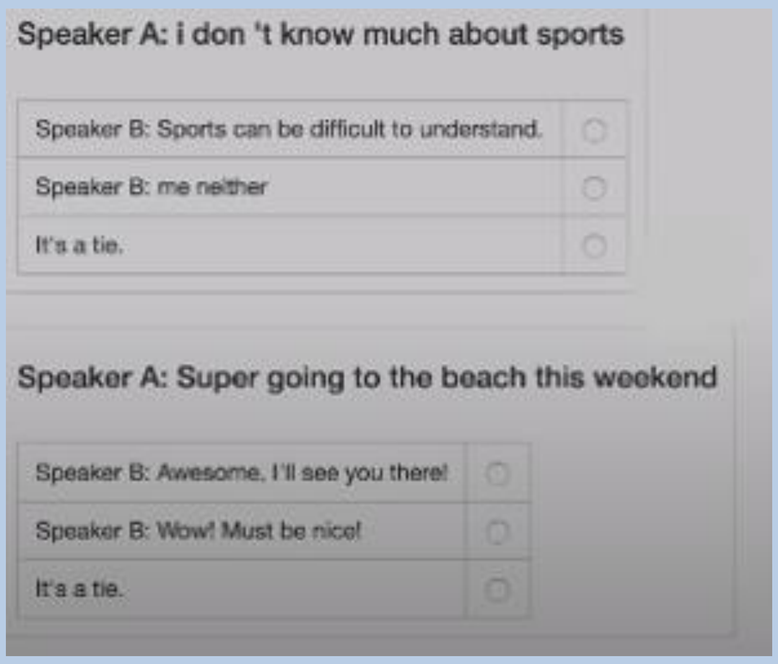


Figure 3: AMT Response [5]

	comparison	graded mean	graded std
0	Cakechat-Seq2SeqAttn_Twitter	-0.604219	0.298198
1	Cakechat-OpenNMT_Seq2SeqAttn	0.055562	0.300693
2	Seq2SeqAttn_OpenSubtitles-Cakechat	-0.545121	0.296132
3	Seq2SeqAttn_OpenSubtitles_without_PTE-Seq2SeqA...	0.044330	0.319641
4	Seq2SeqAttn_Twitter_without_PTE-Seq2SeqAttn_Tw...	0.343868	0.299123
5	Cakechat-NCM	1.292864	0.346975
6	Human1-Seq2SeqAttn_Twitter	-2.051153	0.302624

Figure 4: IRT Evaluation [5]

For instance, the zeroth comparison between Cakechat and seq2seq model on Twitter dataset; negative-graded means that there is a high probability that the cake chat model is better than seq2seq model, plus the confidence, that is, the standard deviation also possess the same. [5]

Project Milestones:

- Data acquired by DBDC.
- Using Amazon Mechanical Turk (AmazonMtruk) chat eval will poll the result of different models.
- Using Item Response Theory, dataset get evaluated using different models.

References:

[1] Appraise: an Open-Source Toolkit for Manual Evaluation of MT Output Christian Federmann DFKI Language Technology Lab https://www.cfedermann.de/pdf/art-federmann.pdf
[2] FNLP 2014: Lecture 15: Machine translation evalution, architecture, lexical models http://www.inf.ed.ac.uk/teaching/courses/fnlp/lectures/15/
[3] Dialogue Breakdown Detection Based on Nonlinguistic Acoustic Information Motoki Abe, Takashi Tsunakawa, M. Nishimura Published 1 October 2018 Computer Science 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE) https://www.semanticscholar.org/paper/Dialogue-Breakdown-Detection-Based-on-Nonlinguistic-Abe-Tsunakawa/3ccd0500eba0c87b4deace01332bced1b11879a7/figure/0
[4] Interrater reliability: the kappa statistic Mary L. McHugh https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/
[5] https://www.youtube.com/watch?v=36rAoujxLAA Date: 1st December, 2021
[6] Naoki Otani, Toshiaki Nakazawa, Daisuke Kawahara, and Sadao Kurohashi. 2016. Irt-based aggregation model of crowdsourced pairwise comparison for evaluating machine translations.