Understanding Conversations for End User Applications

Balamurali A R

Le Laboratoire d'Informatique Fondamentale de Marseille (LIF)

balamurali.ar@lif.univ-mrs.fr

Jan 5, 2016

Conversations

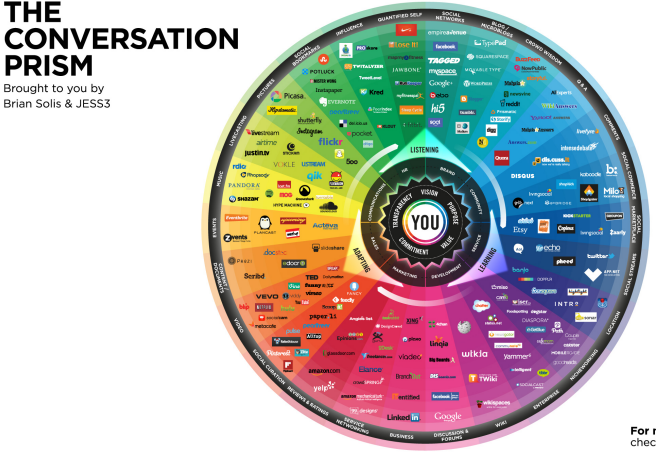

Figure: Different types of conversation happening over Internet

What can we do about it ?

Internet of Things: Network of physical objects or "things" embedded with electronics, software, sensors, and network connectivity, which enables these objects to collect and exchange data

- Information extracted from these conversations act a signals for IoT devices
  - Finding behavior of the person from conversation data to change lighting of the rooms, playlist of music player, cooking suggestion etc
  - Understating intoxicated state to automatically disable the driving mode of the car, sending messages to close friends for pickup, or connecting to cab services

To enable IOT, the focus of the hour is to utilize conversation data for creating end user applications

Overview

1. Analyzing Call Center Operation Data for Quality Monitoring of Customer Calls
2. Analyzing Tweets to Predict Whether User is Alcohol Intoxicated or Not
3. Analyzing Newspaper Comments for Clustering and Labelling

Analyzing Call Center Operation Data for Quality Monitoring of Customer Calls

Part 1: Introduction

- Quality Monitoring (QM) of customer calls is critical for companies[15, 22]
- Performed through call monitoring questionnaires specific to the account [12]
- Random sample of the calls monitored for different dimensions as specified in QM questionnaires by a quality supervisors

However, manual QM is not an efficient process

- An account gets 200-4500 call per day,2 to 10 calls per agent per month is monitored[9]
- Exists a supervisor bias
- QM questionnaire contains multiple dimensions for specific questions, orthogonal assessments are possible
- Attrition rate of employees
- New laws on user privacy mandate strict data masking and storing procedures

Part 1: Objective

Semi-automatic evaluation of the QM for call center data by flagging the conversations which might require deeper inspection by QM supervisors

Part 1: Related Work

They can be broadly classified into three categories:

1. Studies related to queuing, staffing and prediction [17, 20, 21, 18]
2. Studies focusing on improving technological and social environment of the call centers[19, 18, 7]
3. Studies related to behavioral and conversational analysis. [12, 5, 15, 22]

Our study focuses on the latter.

- [12] introduces a quality management system to asses the service of call centers. The system uses a set of features to classify each call as good or bad based on different aspects of the quality questionnaire
- Studies suggest that a customer service can vastly be improved by studying the customer behavior [15]
- When exceptional number of calls being handled, contrasting behavior are seen [1]

Part 1: Methodology

- Annotate a call center data(Decoda corpus) based on QM parameters as PASS or NON PASS as per specific QM parameter
- Create a supervised flagging based system to flag NON-PASS samples for QM supervisors to monitor

Part 1: Quality Parameters

ID	Quality Monitoring Parameters
1	Agent respects opening procedure
2	Agent listens actively and asks relevant questions
3	Agent shows the information in a clear, comprehensive and essential way
4	Agent manages the objections reassuring the customer and always focusing on client satisfaction
5	Agent manages the call with safety
6	Agent uses positive words
7	Agent follows the closing script
8	Agent is polite and proactive with the customer
9	Agent is able to adapt to the style of client's communication always maintaining professionalism
10	Agent Management: he negotiates the wait always giving reasons
11	Ability to listen

Table: Quality Monitoring Parameters Evaluated

Part 1: Annotation Agreement Details

Annotation Categories

- The category PASS reflect that annotator is satisfied with specific objective mentioned in the QM questionnaire.
- If they are unsatisfactory, then they are marked as FAIL.
- If the annotators do not have sufficient information to make decision they are marked as NA.

For developing the Flagging based system, FAIL and NA class together constitute the NON-PASS class

Part 1: Annotation Agreement

Quest. ID	Pass	Fail	NA	Kappa	Agreement
1	346	0	2	1.0	Perfect
2	303	15	30	0.08	Slight
3	334	6	8	0.44	Moderate
4	240	25	83	0.15	Slight
5	329	15	4	0.51	Moderate
6	182	76	90	-0.13	No
7	332	3	13	0.54	Moderate
8	341	4	3	0.65	Moderate
9	330	14	4	0.53	Moderate
10	206	3	139	0.90	Perfect
11	326	16	6	0.25	Fair

Table: Inter-annotation agreement using Fleiss Kappa along with category selected based on majority voting

Part 1: Annotation Agreement Discussion

Moderate inter-annotation agreement.

- QM task is highly subjective (for instance question 2, 4 and 6)
- Multiple dimensions assessed (for instance question 3)

Fully automatic quality monitoring cannot be performed.

- The agreement at the question level is moderate
- QM supervisor differ on the *FAIL* parameters given to justify their decision

Part 1: Flagging based system for Quality Monitoring

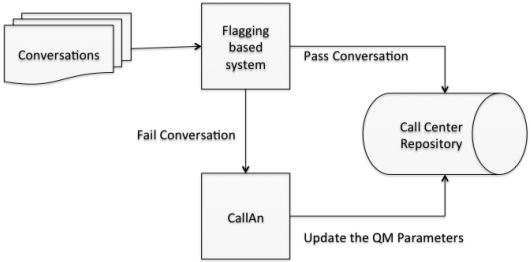

Figure: Process Flow diagram for Flagging based system and CallAn interface

SVM based learner using a linear kernel optimized for maximizing recall [13].

Part 1: Features

- Text Features (TF)
  - Sentiment Features
  - Agent Utterance Features
- Meta Features (MF)
  - Conversation Time Features
  - Conversation Length Features
  - Wait Time Features
- Speech Features (SF)
  - Fundamental frequency,
  - Voicing probability
  - Loudness contour.

Part 1: Flagging based system

Feature Set	PASS			Non-PASS			Overall		
	Precision	Recall	Fscore	Precision	Recall	Fscore	Precision	Recall	Fscore
SF	0.75	0.75	0.75	0.23	0.23	0.23	0.63	0.63	0.63
TF	0.94	0.94	0.94	0.8	0.82	0.81	0.91	0.91	**0.91**
MF	0.75	0.73	0.74	0.23	0.25	0.24	0.63	0.62	0.62
SF+MF	0.75	0.75	0.75	0.23	0.23	0.23	0.63	0.62	0.62
SF+TF	0.79	0.82	0.8	0.36	0.32	0.34	0.68	0.7	0.69
TF+MF	0.94	0.94	0.94	0.8	0.82	0.81	0.91	0.91	**0.91**
SF+MF+TF	0.79	0.82	0.81	0.37	0.32	0.34	0.69	0.7	0.69

Table: Results of Flagging based QM system on various feature sets

Part 1: Analyzing output of Flagging based system

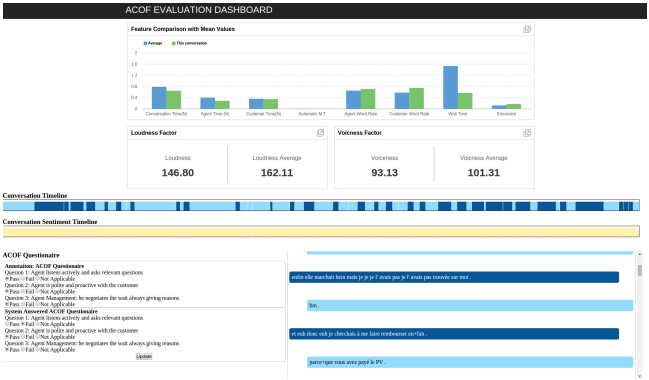

Figure: CallAn User Interface

Code:`https://gitlab.lif.univ-mrs.fr/balamurali.ar/acof-dashboard`

Part 1: Conclusion

- Flagging based approach for call center quality monitoring is proposed
  - High recall based system could be developed to detect NON-PASS call conversation

Future work

- Since Text based features give the best performance, can we use learn a shared representation for two languages
  - Reduces the need for collecting other set of features
- How well does it perform on noisy ASR outputs

Part 2

Analyzing Tweets to Predict Whether User is Alcohol Intoxicated or Not

Part 2: Introduction

- '*I m anakin skywrtr er*'
- '*Retrnd frm wrk l8 2day, d trafic ws bad*'

Which one is written in an intoxicated state?

Part 2: Introduction

- '*I m anakin skywrtr er*' - **Durnk**
- '*Retrnd frm wrk l8 2day, d trafic ws bad*' - **Sober**

Part 2: Motivation

- Past studies show the relation between alcohol use and undesirable social behaviour like aggression [4], crime [6], suicide attempts [11], drunk driving [10] and risky sexual behaviour [3].
- Drunk texting may also be regrettable
  - You must be knowing why
- Driving while intoxicated
  - Highest cause of unnatural deaths in many countries

Part 2: Challenges

1. **More than topic categorization**: [2] show that alcohol users have more pronounced emotions, specifically, anger. In this respect, drunk texting prediction lies at the confluence of topic categorization and emotion classification.
2. **Identification of negative examples**: It is difficult to obtain a set of sober tweets, unless given by the author of the tweet. For example, it is ambiguous whether '*I am feeling lonely tonight*' is a drunk tweet. This is similar to sarcasm expressed as exaggeration (say, '*This is the best film ever!*), where context beyond the text needs to be considered.
3. **Precision/Recall trade-off**: It can be very application sensitive

Part 2: Dataset Creation

To create dataset, Hashtag-based distant supervision as in tasks like emotion classification [14] is used.

1. **Dataset 1**: Collected tweets that are marked as drunk and sober using hashtags. Tweets containing hashtags #drunk, #drank and #imdrunk are considered as drunk tweets, while those with #notdrunk, #imnotdrunk and #sober are sober tweets.
2. **Dataset 2**: Drunk tweets are downloaded using drunk hashtags as above. The list of users who created these tweets is extracted. For the negative class, we download tweets by these users, which do not contain the hashtags corresponding to drunk tweets.
3. **Dataset H**: Drunk tweets are downloaded using drunk hashtags as above. The set of sober tweets is collected using both approaches above. The resultant is the held-out test set Dataset-H which contains no tweets in common with Datasets 1 and 2.

Part 2: Features

N-gram Features			
Unigram (Presence)	Bigram (Presence)	Unigram (Count)	Bigram (Count)
Stylistic Features			
LDA unigrams (Presence)	POS Ratio	#Named Entity Mentions	#Discourse Connectors
LDA unigrams (Count)	Spelling error	Repeated characters	Capitalization
Length	Emoticon (Count)	Emoticon (Presence)	Sentiment Ratio

Table: Feature set for our drunk text prediction system

Part 2: Dataset Creation

To create dataset, Hashtag-based distant supervision as in tasks like emotion classification [14] is used.

1. **Dataset 1**: Collected tweets that are marked as drunk and sober using hashtags. Tweets containing hashtags #drunk, #drank and #imdrunk are considered as drunk tweets, while those with #notdrunk, #imnotdrunk and #sober are sober tweets.
2. **Dataset 2**: Drunk tweets are downloaded using drunk hashtags as above. The list of users who created these tweets is extracted. For the negative class, we download tweets by these users, which do not contain the hashtags corresponding to drunk tweets.
3. **Dataset H**: Drunk tweets are downloaded using drunk hashtags as above. The set of sober tweets is collected using both approaches above. The resultant is the held-out test set Dataset-H which contains no tweets in common with Datasets 1 and 2.

Part 2: Classification Results

	A (%)	NP (%)	PP (%)	NR (%)	PR (%)
Dataset 1					
N-gram	85.5	72.8	88.8	63.4	92.5
Stylistic	75.6	32.5	76.2	3.2	98.6
All	85.4	71.9	89.1	64.6	91.9
Dataset 2					
N-gram	77.9	82.3	65.5	87.2	56.5
Stylistic	70.3	70.8	56.7	97.9	6.01
All	78.1	82.6	65.3	86.9	57.5

Table: Performance of our features on Datasets 1 and 2; A: Accuracy, PP/NP: Positive/Negative Precision, PR/NR: Positive/Negative Recall

Part 2: How does Humans fare on detecting drunk state

Annotation agreement on Dataset H (Held out data)

	A1	A2	A3
A1	-	0.42	0.36
A2	0.42	-	0.30
A3	0.36	0.30	-

Table: Cohen's Kappa for three annotators (A1-A3)

Part 2: Held out data evaluation

	A (%)	**NP (%)**	**PP (%)**	**NR (%)**	**PR (%)**
Annotators	68.8	71.7	61.7	83.9	43.5
Training Dataset	**Our classifiers**				
Dataset 1	47.3	70	40	26	81
Dataset 2	64	70	53	72	50

Table: Performance of human evaluators and our classifiers (trained on all features), for Dataset-H as test set; A: Accuracy, PP/NP: Positive/Negative Precision, PR/NR: Positive/Negative Recall

Part 2: Summary

- Introduced the task of drunk text prediction
- Possible to predict drunk state by analyzing tweets

Future work

- Capturing the keystrokes from Android keyboard and modelling the sober state
- Detecting the drunk state from keystrokes and analyzing the actual text content

Analyzing Newspaper Comments for Clustering and Labelling

Part 3: Demo

Demo :http://pageperso.lif.univ-mrs.fr/~balamurali.ar/sensei.html
http://139.124.5.125/vis/examples/network/interactive.php?article=281

Part 3: Introduction

- Online news outlets attract large volumes of comments every day. *The Huffington Post*, for example, received an estimated 140,000 comments in a 3 day period, while *The Guardian* has reported receiving 25,000 to 40,000 comments per day

How to assimilate and comprehend them?

Part 3: Solution

Automatically generate end-user friendly topic clusters of reader comments to online news articles. We propose graph-based methods to address two tasks:

1. To group reader comments into topic clusters
2. To label the clusters for the topic(s) they represent

Part 3: Solution

Automatically generate end-user friendly topic clusters of reader comments to online news articles. We propose graph-based methods to address two tasks:

1. To group reader comments into topic clusters
2. **To label the clusters for the topic(s) they represent**

Part 3: Labeling

Topic labelling algorithms are extractive[16]

- *red, blue, green, yellow* $\Rightarrow$ *red*

However, abstractive labeling makes life easier

- *red, blue, green, yellow* $\Rightarrow$ *type of colors*

Part 3: Abstractive Labelling

Modify the graph-based topic labelling algorithm described in [8] Basically the algorithm uses

- Topic graph of all the topics that the cluster represents
- Use DBpedia to create graphs
- Expand using DBpedia relations
- Use centrality measure to find the common label that encompasses

Part 3: Abstractive Labelling

Require: topicsList(θ), relations to expand (X), graph centrality measure(R)

1: $label \leftarrow empty$
2: **for all** $\theta_j \in \theta$ **do**
3: $\quad T^{\theta_j} \leftarrow initializeGraph(V, E, \theta_j)$
4: $\quad$ **for all** $W_i \in C^{\theta_j}$ **do**
5: $\quad\quad G_i \leftarrow expandGraph(X, T^{\theta_j}, W_i)$
6: $\quad\quad T^{\theta_j} \leftarrow G_i \cup T^{\theta_j}$
7: $\quad$ **end for**
8: $\quad label^j \leftarrow getCentralNode(T^{\theta_j, R})$
9: $\quad label \leftarrow label \cup label^j$
10: **end for**
11: **return** label

Part 3: Generated Label

I was working in the south of France during the 2003 heatwave and I would wear a hat which I kept soaked in water while at work, lie in a bath full of cold water before going to bed to get my body temperature down and drink gallons of water through out the day and before going to bed. My thermometer only went up to 50 degrees centigrade so I don't know how hot it was during the day but at 3 in the morning it was 32 degrees centigrade. I also avoided alcohol. It was hot. Here in Devon it's rarely been under 75 since June. In these conditions, dogs die due to stupid owners leaving them in cars, people under-hydrate and pass out, wild-fires start due to fag ends on dry grass, all sorts of stuff happens. It depends where you are, but some of the UK is baking.....

System Generated Label:**Occupational safety and health**

Part 3: Labeling Evaluation

Qualitative Evaluation of Labels

- Create system generated labels
- Give the comments to choose from
- Choice provided for no labelling scenario as well

Calculate the accuracy and the inter-annotation agreement

Part 3: Annotation agreement

Annotators	A-B	B-C	C-A	Overall
Agreement	0.76	0.45	0.64	0.61

Table: Annotator agreement (Fleiss Kappa) for comment labelling over 22 comment clusters

Part 3: Annotation accuracy

Annotator Data	Precision	Recall	F-score	NA
A	0.7	0.59	0.61	12
B	0.73	0.62	0.61	11
C	0.85	0.76	0.75	2
mean	0.76	0.6566	0.66	

Table: Evaluation results of the cluster labeling system for each of the 3 annotators

Part 3: Summary

- Introduced a graph based abstractive algorithm
- Sensible label creation - However, not perfect

Future Work

- Create facts to support the labels generated

Code

- `https://gitlab.lif.univ-mrs.fr/balamurali.ar/topic_labeller`
- `https://gitlab.lif.univ-mrs.fr/balamurali.ar/ISummerization`

Acknowledgment

Aditya Joshi, Abhijeet Mishra, Jeremy Trione, Ahmet Akter, Emina Kurtic, Monica Paramita, Mickael Rouvier, Benoit Favre, Fredric Bechet


Q&A

Reference

Salah Aguir, Fikri Karaesmen, O Zeynep Akşin, and Fabrice Chauvet.
The impact of retrials on call center performance.
OR Spectrum, 26(3):353–376, 2004.

Josephine A Borrill, Bernard K Rosen, and Angela B Summerfield.
The influence of alcohol on judgement of facial expressions of emotion.
British Journal of Medical Psychology, 1987.

Angela Bryan, Courtney A Rocheleau, Reuben N Robbins, and Kent E Hutchinson.
Condom use among high-risk adolescents: testing the influence of alcohol use on the relationship of cognitive correlates of behavior.
Health Psychology, 24(2):133, 2005.

Brad J Bushman and Harris M Cooper.
Effects of alcohol on human aggression: An intergrative research review.
Psychological bulletin, 107(3):341, 1990.