Comcast Confidential - Any unauthorized disclosure or use is strictly prohibited

OpenNLP Research

Using Context to inform OpenNLP Research - DEC 2019

Identification of Orphan Domains

Child Pages

We still have lots of N/A queries queries which are outside of known boundaries and not handled by any agent in Corti-NLP. We're trying to handle as much these queries as possible by addressing new/orphan domains. We identified some new orphan domains by investigating current failed queries. Also, some app-specific orphan domains are introduced to enhance the user experience of Xfinity app such as YouTube, Pandora, etc.

Info

What is a Domain? What's the difference between agent and domain?

https://github.comcast.com/pages/compass-vrex/corti-docs/#/onlp/overview

Challenges

1. N/A queries related to contents (especially from sources outside like YouTube, Netflix, etc) are hard to identify without additional information (Note: We try using external APIs to collect more information for these queries in another project)
2. Most of these N/A queries are short with a few tokens, which makes it hard to capture semantics
3. Some queries include ASR errors, which sometimes distorts or makes it lose original semantics

Comcast Confidential - Any unauthorized disclosure or use is strictly prohibited

Annotations (for Orphan Domains)

Based on the investigation of NA queries, we've identified some new orphan domains (Refer to Real query investigations for unhandled domains and gibberish).

Domains

- Contents
- Apps
- Knowledge
- Recommendation
- TV Guidance
- Gibberish (Abbreviation / Ambiguous / Incomplete / SR Issue / Repetition)
- Device control
- Channels

Here is the list of the annotated samples (i.e. with new orphan domain labels) to evaluate the performance of the model.

1	https://comcastcorp.sharepoint.com/:x:/r/sites/AppliedAI/Shared%20Documents/Projects/Corti/domains/opennlp/annotation/annotation_new_opennlp_domains-no_domain_samples_revised.xlsx?d=w8c2b97c88e7441e99eb02825c1995fbf&csf=1&web=1&e=6GwM2F
2	https://comcastcorp.sharepoint.com/:x:/r/sites/AppliedAI/Shared%20Documents/Projects/Corti/domains/opennlp/annotation/annotation_new_opennlp_domains-no_domain_samples2_revised.xlsx?d=wc6d9f3281f5c45b9a786d153ae3beef5&csf=1&web=1&e=mQjPT
3	https://comcastcorp.sharepoint.com/:x:/r/sites/AppliedAI/Shared%20Documents/Projects/Corti/domains/opennlp/annotation/annotation_new_opennlp_domains-no_domain_samples3.xlsx?d=w0895c3bd40604b6e8a248f1404d55b94&csf=1&web=1&e=qwBfGy
4	https://comcastcorp.sharepoint.com/:x:/r/sites/AppliedAI/Shared%20Documents/Projects/Corti/domains/opennlp/annotation/annotation_new_opennlp_domains-no_domain_samples4.xlsx?d=wd42be763b5c7421ba742f5c78bc18d8c&csf=1&web=1&e=Ofn8ta

Comcast Confidential - Any unauthorized disclosure or use is strictly prohibited

Annotations (for Gibberish Domains)

We observe lots of gibberish (or gibberish-like) queries every day, which degrades the quality of systems and increased the traffic of NLP systems. However, the automatic detection of these queries are really challenging as the patterns inside queries are not clear.

Our first step for detecting these queries effectively is to do manual investigation against some target queries.

In particular, we focus on (1) high frequency queries and (2) queries with negative user interactions, to maximize business impact and reduce the scope.

- New field "Rank" is introduced for the purpose of filtering (9 or higher), which can be calculated as follows: tot frequency * (neg_ratio - pos_ratio) / SQRT(len(query))

We also addressed a few new fields to guide annotators in their investigation task like following.

- SortedSuggestions & TopSuggestion: Suggestions based on Serene Prediction
- DomainsToRun: If YouTube is captured, this query might be related to YouTube domain

Labels

Label	V1	V2 (Anthony's version)
Suggestion	Provide "suggested" query for suggestion (Assume we have one clear answer)	Provide "suggested" queries for suggestion (Regardless of number of suggestions)
Clarification	Ask the user which suggestion to choose when there are multiple candidates	Don't rely on suggested queries. Suggest current one or new query
Gibberish	Gibberish	Gibberish
Neutral	Keep current action	Keep current action

Annotation Result

Original SpreadSheet:https://comcastcorp.sharepoint.com/:x:/r/sites/AppliedAI/_layouts/15/Doc.aspx?sourcedoc=%7B0CF91941-6BC9-443E-A12C-CE1864DD72BD%7D&file=negatives_with_suggestions.xls&action=default&mobileredirect=true

Annotated SpreadSheet: https://comcastcorp.sharepoint.com/:x:/r/sites/AppliedAI/_layouts/15/Doc.aspx?sourcedoc=%7BFF6CCE13-76F6-4B0B-AA43-D5C57654A2A5%7D&file=negatives_with_suggestions_v2.xlsx&action=default&mobileredirect=true

Comcast Confidential - Any unauthorized disclosure or use is strictly prohibited

Clustering approach (I) to identify Orphan Domains

Approach

The most common approach to identify/classify utterances in an unsupervised way is clustering method. We've tried building clusters for N/A queries that we collected in APRIL (some of them are annotated) by using K-means clustering algorithm.

Clustering with BERT embeddings

We've tried K-Means clustering method to get clusters (cluster size = 25). As K-Means requires a normalized vector of each query as input, we've applied pre-trained BERT model (uncased, 12 layer, 768 size) to get sentence embeddings for all target queries. We initially tried 10 clusters for clustering and noticed the granularity was too coarse to come up with clusters representing unique characteristics. More trials with different cluster size can be done in the future.

Evaluation

Since we have some N/A queries annotated, we can compare the results of these clusters with DOMAINS that annotator identified to see how meaningful these cluster are.

Experiment Result

Cluster Analysis (with 25 clusters)

Here are the characteristics of each cluster that we've identified by looking at main patterns in the cluster. For more details (also all raw data), you can check this spreadsheet: https://comcastcorp.sharepoint.com/:x:/r/sites/AppliedAI/Shared%20Documents/Projects/Corti/domains/opennlp/research%20docs/cluster_analysis.xlsx?d=w6c4fcb60557549ebad3e72c65ffa927c&csf=1&web=1&e=3Q8NVA

Cluster	Domains/Patterns	Samples
1	GENERAL DEVICE CONTROL	audio settings, check battery, password, switch off
2	TIME	fast forward three minutes 32nd, set timer for 10 minutes, skip two minutes, 60 minutes please, 2: 00 PM
3	SPORTS INFO	nba league pass, minnesota state hockey tournament, austin celtics
4	CONTENTS	tariq l moussa, five nights at freddy's, bonfire of the vanities, strayhan sara and keke
5	MUSIC	rock 'n' roll hall of fame, children's songs, pj masks song, baby shark remix, free blues music
6	FOREIGN CONTENTS	vivir del cuento, gordo y la flaca, la chona, concierto de juan gabriel
7	RECOMMENDATION	free movie wake, movies with the rock, shows for toddlers, free movies on tonight, free movies for teenagers, lego movie free
8	GIBBERISH (Very short), KNOWLEDGE (Location)	cd, pg, mr. pd, aftica, australia, hq, ba, av, ask, update, france, vc, sv, thailand
9	"The -" pattern & GIBBERISH	the chrisleys, the bluebook, the sims, the habs and the have, are the, the 72
10	GIBBERISH with Repetition, "greeting" pattern	good night, ha ha, no no no, yo yo yo, blah blah blah blah blah blah blah, hello man, no go
11	NAME, "bill" pattern	dave turin, bill mauer, gabby durant, lacy peterson, savage bill, brandon frazier, dave fane, stimulus bill
12	QUESTION	so you think you can dance, what's, what's on tv tonight, is there, how many, how old are you
13	Abbreviation (CHANNEL/GIBBERISH)	are fd, o wn, l, s ec, la ff, blu e, tw, wt, hl, l a ff, de vs, wg, xm, ew
14	APPS, DEVICE CONTROL	amazon plus, video descriptive off, download disney+, xbox on, show me wi-fi, bluetooth on, dvr subscription, add hulu app
15	"i -" pattern	i owe n, i need to, i want to see, i still believe movie, i'm not okay with it, i miss, give me the
16	CHANNEL+	cartoon net, hallmark movies and missed, children's tv, laugh network, tv shows on tonight, dr. phil on cbs, record yesterday on hbo
17	CONTENTS (Popular Search Keywords)	corona virus, epix free, delphia flyers, disney songs, ozark trailer, barbie games, echo canyon
18	NUMBER+ (CHANNEL)	45 tv, ppv, oj 25, cw 50, pn two, p hl 17, 19 please, kc 24, 30 please, hi 01

Comcast Confidential - Any unauthorized disclosure or use is strictly prohibited

19	CONTENTS (Movie)	spider-man into the spider, scooby-doo: the movie, the rocketman, the movie mr. rogers, dr. seuss's movie
20	Xfinity	xfinity password, xfinity on, xfinity information, xfinity specials, xfinity games, xfinity mobile, xfinity upgrade
21	APPS (Neflix/Amazon)	netflix mark wahlberg, download netflix, netflix streaming, amazon prime bosch season 2, netflix thompson girl
22	CHANNEL	at&t, hbo on, nfhs network, fx fm, fox net, s ec network, epix on, gh tv, channel 4 cbs, bb&t, de tv, vpn
23	SPORTS INFO (Possesive case)	us women's soccer team, men's michigan basketball, byu vs. st. mary's, ncaa men's
24	CONTENTS with Number (two, three, ..)	trolls two, minions two, boo two, mulan two, pc three, thor two, disney two, cbs two, hellboy two
25	CONTENTS with Command, Verb	watch the winsors, is loveblind, play the monster machine, have and have, watch the lorax, play hip-hop miami

M/L model for YouTube domain detection

Model Building/Training

The goal of this research is to build M/L model that detects the queries belong to YouTube domain (non-official at this point). We tried rule-based model prior to this approach but rule-based approach had a few caveats like this:

1) hard to deal with variations (that includes mis-spelling) of content titles / channel titles

2) hard to detect gibberish by rules

Feature definition

Based on outputs of YouTube Data API and some insights from initial investigations, we've identified following input features

Feature	Note
query	given utterance
popularity	popularity of query (obtained by YouTube API)
channel	name of the YouTube channel if exist
playlist	name of the YouTube playlist if exist
contents 1 - title	title of the first YouTube contents
contents 1 - channel	channel for the first YouTube contents
contents 1 - time published	time published for the first YouTube contents
...	
contents 5 - title	title of the fifth YouTube contents
contents 5 - channel	channel for the fifth YouTube contents
contents 5 - time published	time published for the fifth YouTube contents

Label

Binary classification: 1 if YouTuble domain, 0 if not

Model

- Use Character-based CNN model to get embeddings for <query, channel, playlist, title for each content, channel for each content>
- Apply dot product, subtraction, maximum of (query, channel/playlist/content title/content channel) to get the similarity signals
- ~~Also considers similarity between contents (Dropped in the middle)~~
- Choose the maximum similarity among (query-channel similarity, query-playlist similarity, query-contents similarity)
- Consider adjustments based on popularity and time published

Data preparation

Annotated 2,569 samples by looking at YouTube Data API result for target sample queries.

Data balancing

Due to unbalanced data set (i.e. number of positives is much higher that number of negatives), we've utilized a generator to build each batch data where part of positives samples are used.

Training/Prediction result

Depending on situations, model sometimes didn't converge well. Sometimes (after slight change in model), the model got converged to some point but the accuracy still arrived at around 80%, which is not quite different from the accuracy of dummy model that predicts only positive.

Analysis

Challenges in data (annotation)

There are a lot of ambiguities in decision making due to

1) existence of gibberish, which mostly returns high score (e.g. "when are")

2) existence of mis-spelled queries (e.g. "blippy" instead of "blippi")

3) some queries are specific to one content or channel name or playlist title, while others are very generic queries that include all searched contents (e.g. "peppa pig" VS "cute things")

Limitation of Character-based CNN (with small number of data)

As models rely too much on characters that do not have semantics inside, it's hard to generalize the model to unseen patterns. Also, small number of models make the model susceptible to variations of language patterns.

Possible improvements in the future

1. The model will be better with more data
2. Combination with Rule-based model
3. Concentrate on some specific cases like query-channel/playlist/contents_title similarity, while solve other problems (e.g. Gibberish) with other models

Comcast Confidential - Any unauthorized disclosure or use is strictly prohibited

Clustering Approach (II) to identify General Knowledge domain

As we are collaborating with Microsoft to handle queries that request for answer from knowledge sources, we need a classifier/detector for these queries to forward them to Microsoft Knowledge agent. Here, we're going to do POC work of knowledge query identification to ensure its feasibility.

Challenges

- How can we capture "General Knowledge" domain without annotated data
- How to make a distinction between "Comcast-specific questions or FAQ queries" and "Generic knowledge questions"

Approach

Data collection

Comcast Queries (100K) + Google Natural Questions (40K) + Kaggle Local Search Queries (30K)

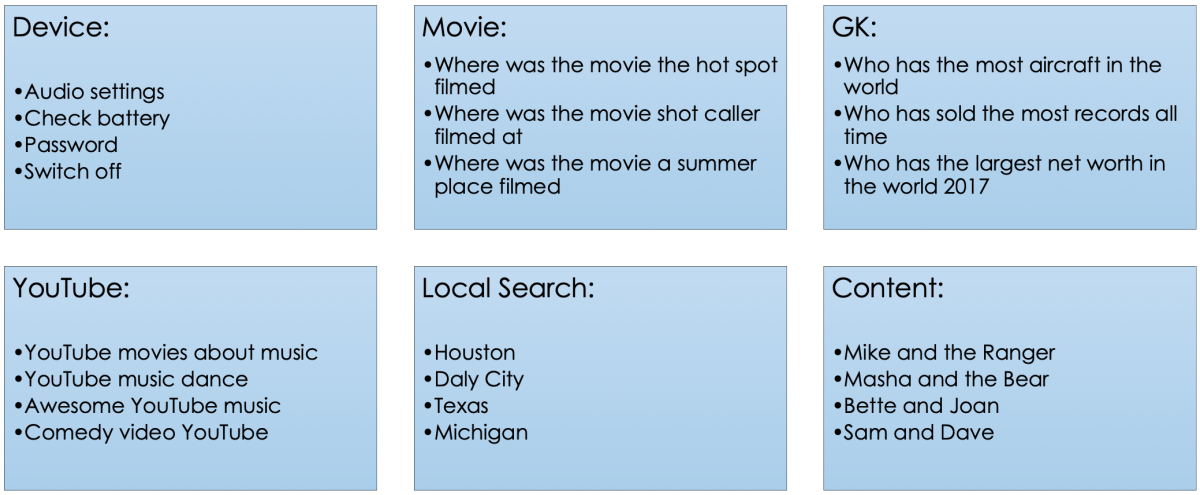
Clustering Approach

Refer to : https://comcastcorp.sharepoint.com/:p:/r/sites/AppliedAI/Shared%20Documents/Projects/Corti/domains/opennlp/research%20docs/Report_PPT_POmkar.pptx?d=w238fa04c4a0640c8a0e6a61c9719b577&csf=1&web=1&e=B3E0fe

1. Get vectors for the queries
2. Build clusters by using K-Means clustering algorithm
3. Observer patterns in each cluster and find clusters related to "General Knowledge"
4. Label queries based on identified clusters
5. Train new classifier based on labeled data

Result

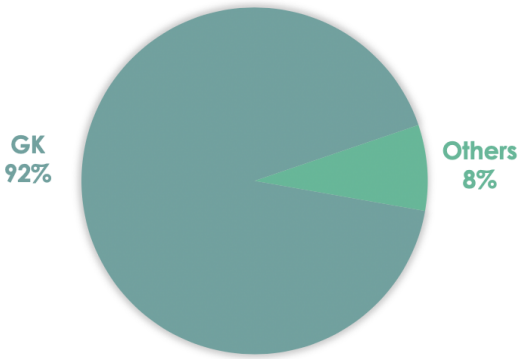
- Obtained 150 Clusters
- Observed patterns in some clusters like following



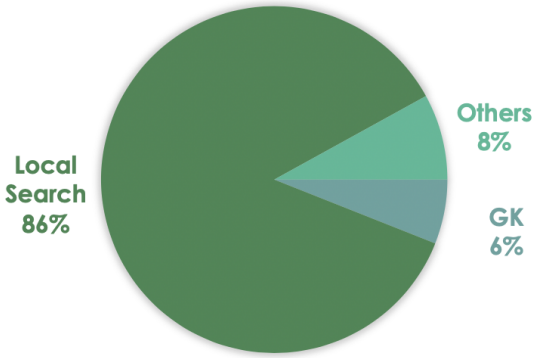
Evaluation

Used 300 GK & Local Search queries for the evaluation of final classifier.

GK



LOCAL SEARCH



Detection of Gibberish Domain Queries

Analysis of current gibberish queries

For the purpose of gibberish query analysis, we've extracted some useful information from session data (AUG 29) like below.

Column name	Description
query	target gibberish query (see attached "gibberish-queries.jsonl")
count	occurrences of target gibberish query (per 1 day)
query_repeat (query event repeat)	number of query event repetition (QUERY event right after target query without any other event) (e.g.) QUERY (target gibberish query) - QUERY
negative_event	number of target query sessions containing any negative event (EXIT, MENU, GUIDE, SEARCH) • consider only first 4 events after query (e.g.) QUERY - KEY_EXIT
next query - top n	next query: query that was tried after target query top n: the n-most frequent next query

- Target gibberish queries

You can check the analysis document,

https://comcastcorp.sharepoint.com/:x:/r/sites/AppliedAI/Shared%20Documents/Projects/Corti/alf/orphan%20domains/gibberish_analysis.xlsx?d=w3a069b2c6d2e42569882cfea04d8decd&csf=1&web=1&e=bABjo3

with columns as shown in the table.

Detection of Music Domain Queries

Comcast Confidential - Any unauthorized disclosure or use is strictly prohibited

Evaluation of Music Domain Detection

Evaluation Result

	POSITIVE QUERIES	NEGATIVE QUERIES
New Version in ALF Pipeline	33 Positives / 65	**35 Positives / 35**

Evaluation Set

POSITIVE QUERIES (65)	NEGATIVE QUERIES (35)

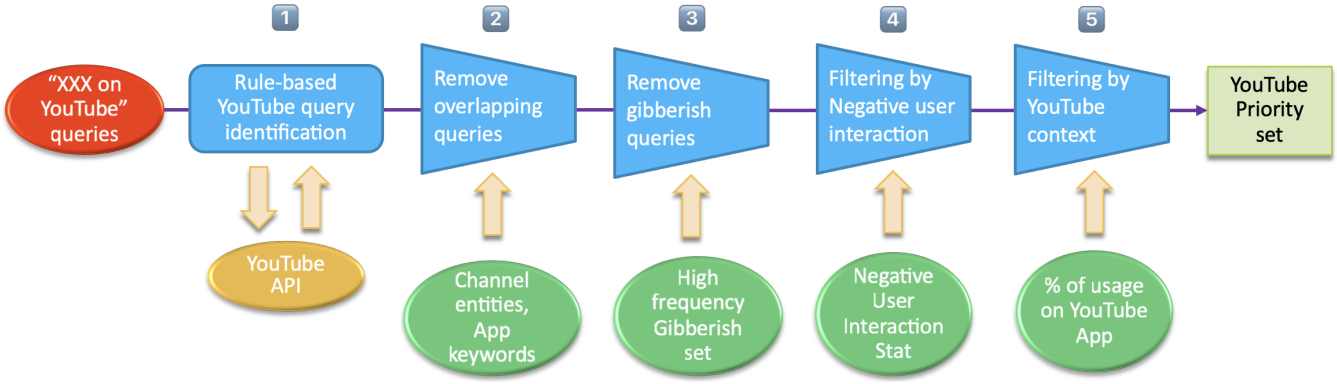
Phil Collins in the air tonight,1 Little Lies by Fleetwood Mac,1 Diana Krall,1 Hardcastle,1 Guns and Roses,1 Neil Diamond,1 Beastie Boys,1 Suzy Bogguss music,1 Imagine dragons,1 Jason Aldean,1 Beatles,1 Justin Bieber,1 Mandela effect,1 Sting,1 Biggie Smalls,1 Tom Petty music,1 Bobby Darin,1 Kehlani,1 Booker T and the MGs,1 Supertramp,1 Dua Lipa,1 Motown,1 Linkin Park,1 Old Dominion,1 the Manhattans,1 DNCE,1 Death cab for cutie,1 J Boog,1 Life Jennings,1 Play music by Ludacris,1 Dusk till Dawn,1 Die happy man song,1 Song halo,1 Be happy,1 I hope,1 Come and go,1 fireball,1 Baby Shark,1 Kidz bop,1 Star Trek Voyager,1 Moana soundtrack,1 I heart Christmas music,1 Army Wives,1 Easy listening,1 Hawaiian Music,1 Today's hip-hop and R&B,1 Praise and Worship Music,1 Soft piano music,1 Baby Sleep music,1 Kids Disney,1 trance instrumentals,1 Jazz music,1 70s radio,1 Brantley Gilbert radio,1 children's songs,1 kid rock,1 Light jazz radio,1 London grammar radio,1 Heavenly lullabies,1 hipster cocktail party,1 90s country hits,1 Rick Ross radio,1 Oldies but goodies,1 thumbprint,1 uplifting music,1	cab Mo,0 Fluffy unicorns on put,0 Hood Rich Taylor,0 Ferruccio by the Glades,0 some Danny,0 Joe Biden,0 Corona Virus,0 HDMI,0 Input Password,0 What is the woodpecker,0 Corded programs,0 Showtime After Dark,0 Christmas movies for children,0 Skip commercial,0 NFL Sunday night game,0 The Croods a New Age,0 CBS 24,0 Inquisitor master,0 Foods,0 Recently deleted,0 reprendre au début,0 Tom Hanks Greyhound,0 Governor Cislak,0 Back one minute,0 Fast forward 30 seconds,0 Master flippers,0 ABC app,0 A sugar and spice holiday,0 Record the AMA's,0 TV brightness,0 Queue the Sea,0 Volume settings,0 NJ news,0 Rain sound,0 The Siesta dress,0

Detection of YouTube Domain queries

Child Pages

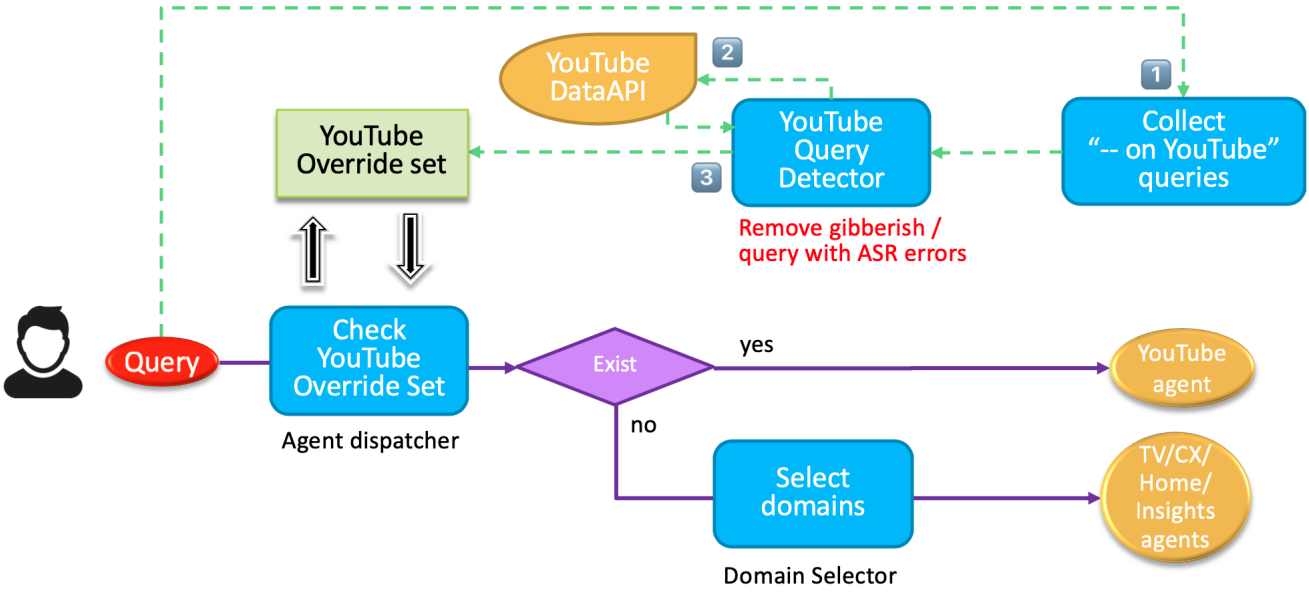
- Evaluation of YouTube Query Detection
- Investigation of YouTube domain queries detected to remove Gibberish
- Investigation of YouTube domain queries with many neutral (or mixed) user feedbacks
- Refine rules for YouTube detection

Identification steps



Step 1: Rule-based YouTube query identification

In order to detect YouTube queries, we first collect potential YouTube queries from previous successful query trials with guard word such as "-- on YouTube". Then we apply rule-based YouTube query detector for additional filtering which utilizes external YouTube data api to make a better decision.



- Detected YouTube queries (for queries tried during last 100 days at 1 peak hour, frequency equal to or higher than 10)) : https://comcastcorp.sharepoint.com/sites/AppliedAI/_layouts/15/download.aspx?UniqueId=5be8287f1d714e58a44e55cf570ddf20&e=cls3d7

Checkpoint

The detected queries still contain gibberishes such as incomplete queries ("is it", "I was", "this is", ...) or queries with short tokens ("o", "e", "wow", ...).

Step 2: Remove station queries

The result from step 1 contains lots of known channel names and app titles which can be captured by TV agent correctly. We remove these overlapping queries to focus on target YouTube domain.

Step 3: Remove gibberish queries

We observed that lots of high frequency queries which were predicted as YouTube domain query turned out to be gibberish. We identified these high frequency gibberish queries by trying both automatic detection and manual reviews in order to filter them out in the result of step 2.

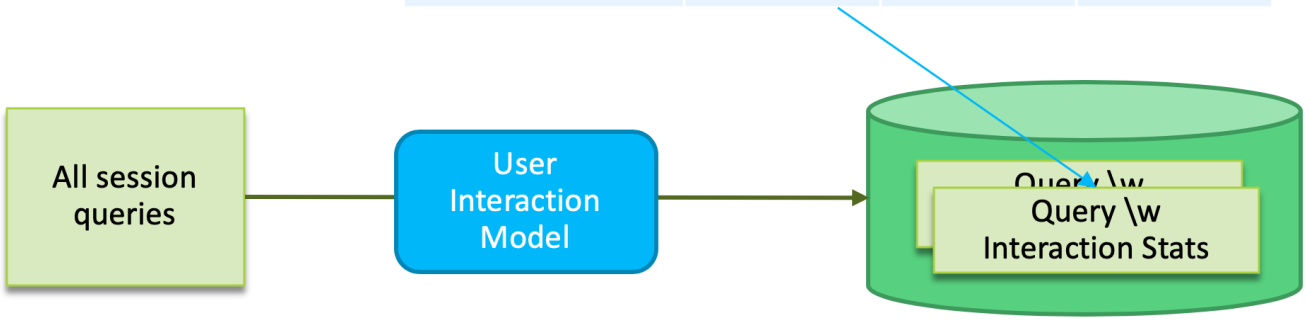
Step 4: Filtering by negative user interactions

As the result of step 1-3 still contains many valid VREX queries with positive user feedbacks, we identify positive examples from the result of Approach 1 in terms of user interaction (by applying addition user interaction model) and remove these from the priority set.

Positive interaction from the user interaction model represents the successful execution of the query such as showing the right contents for the query, while negative interaction represents failed execution such as returning N/A response, showing result not relevant to the query, no response, etc. For example, trying "pokemon" shows the list of contents related to "pokemon", which led to positive user experience. In the case of "baby shark remix", the user experience is negative as the retrieved content is not the one that user expected (although it's related).

The strategy here is to filter out the positive queries from the initial YouTube priority set obtained in the approach 2 since our systems (TV agent) can handle these queries without the information of YouTube domain. However, positive queries from the user interaction model are still valid YouTube contents. (Both Xfinity and YouTube support accessing these contents).

Query	# Positives	# Neutral	# Negatives
pokémon	409	180	77
baby shark remix	0	0	10



- Positive YouTube queries : https://comcastcorp.sharepoint.com/sites/AppliedAI/_layouts/15/download.aspx?UniqueId=7fc294fc11914d72973bb24081118efd&e=16aYic
- Refined YouTube queries (Result of Approach 1 - Positive) : https://comcastcorp.sharepoint.com/sites/AppliedAI/_layouts/15/download.aspx?UniqueId=a1b09f165eee4caca2f42a3fe969eabd&e=DXeOXL

Step 5: Filtering by YouTube context

To be more conservative on the YouTube priority set (i.e. increase precision of the result), we check if queries were executed on YouTube app context. We filter out queries of which usages (percentage) on YouTube app context is lower than given threshold (10%).

Evaluation of YouTube Query Detection

Evaluation Result

	POSITIVE QUERIES	CHALLENGING POSITIVE QUERIES	NEGATIVE QUERIES
Old Version in OpenNLPPipeline	93 Positives / 100	77 Positives / 100	**29 Positives / 57**
New Version in ALF	79 Positives / 100	52 Positives / 100	**13 Positives / 57**

Evaluation Set

POSITIVE QUERIES (100)	CHALLENGING POSITIVE QUERIES (100)	NEGATIVE QUERIES (57)
123 go!	10 minute ab workout	music is
90s hip-hop	prank videos	de son
ImJayStation	fortnite funny moments	how to make
coco mellon baby shark	3d	sbl
a for adley	salsa	vhc
ab workout	queen live aid	add play
america's funniest home videos	chloe ting two-week shred	two songs
anna and elsa	baseball bloopers	versus
asmr	if you're happy and you know it clap your hands	frank
baby lullabies	cookie swirl c playing roblox	currently
baby shark cocomelon	daisy and molly	a song
be amazed	charlie demilio tick-tock	change to kids
beginner yoga	spider-man and hulk	fleming
bff squad	piggy jump scares	i want to go
blippi christmas	tractor videos	yummy yum
blue's clues	bill winston	hi this
bon appétit	ladybugs	clean
bongo cat	spooky scary skeleton	for show me
brianna plays	peekaboo elmo	say show
broken tv prank	middle school wrestling	what does
bts dynamite	peppa pig memes	and beyond
bubble butt	jazz ente	true true
busy beavers	30 days of yoga with adrian	in the middle
carpool karaoke	ryan show	z
cartoon cat	preston and brianna	and we know
charlie demilio tick-tock	hello neighbor two	dell fast
chef peepee	people eating food	do you
chloe ting workout	troll songs	fox say
choo-choo train	crosby stills nash and young	give mouse
cj so cool songs	i believe i can fly	gmm
cookie swirl c playing roblox	hilarious videos	go to
coronavirus	puppy videos	he mouse
cringe fam	the chicken dance	i was
crooked media	prince and the revolution	is it
cub scouts	rock 'n' roll	old are
damien and bianka	baby shark r&b	add me
dave and ava lullabies	fart sounds	84 la
david jeremiah	descendents three songs	history on me
dead and company	how to get free roebucks	what are
deshi games	20th century fox	can you
dinosaur toys	lsu football	been
dirt bikes	little ron ron	a friendly
disney sing-along	race cars for kids	ass mall
don't hug me i'm scared	happy by pharrell	and
down by the bay	sofia the first theme song	i dvr
drag racing	gymnastics	are
e!	vampire	game f
elsia and anya	new skate pro	no no
exercise for kids	how to make lipgloss	i smell
f gtv playing piggy	mike tyson highlights	i am elizabeth
fall guy's	dancing fruit	hello g
fernwood tonight	snowman	what are you
five minute crafts	ceiling fans	to madison
five nights at freddy songs	tinfoil hat	all off

fortnite item shop	spider-man for kids	tap and
free western movies	chad wild clay and v quaint	i am i savage
frozen ii songs	surfing	not
funk brose	magic	
funny dog videos	zumba for beginners	
funny tik tok	cooking videos	
gabby and alex	walk with leslie	
game of zones	make up tutorials	
garbage trucks	mermaids	
golf lessons	trolls world tour soundtrack	
google gaga baby	dinner and a movie	
goon squad	nick a 30	
gotcha life	walking exercise	
gta five	bedtime stories for babies	
guess the emoji	floss tube	
hairstyles	sasha and shiloh	
halloween music	devils don't fly	
has fit	hairstyles	
hey dougie	that girl lele	
how to make slime	super smash bro's	
hurricane laura	marvel racing	
hydro dipping	elmo brush your teeth	
i has cup quake	slither i/o	
itsy-bitsy spider	godzilla vs. king kong	
j house vlogs	nokia	
jelly fortnite	four wheelers	
joe biden speech	ryan's toy review	
joe josie wah	free peppa pig	
johnny johnny yes papa	moana you're welcome	
juice world	it's raining tacos	
jump scares	hip-hop dances	
just dance 2020	meditation music	
kakko	i of the tiger	
karaoke songs	birthday songs	
kc and rachel	kids dancing	
kids bop	roddy rich the box	
kids zumba	remix songs	
labrant family	bear	
lakewood church	dog noises	
lego	try not to cry	
lip sync battle	the chosen episode 5	
little baby bump	toddler cartoons	
lol surprise dolls	unspeakable hide and seek	
toy story four	clean tick-tock mash up	
gma news	mickey mouse theme song	
quarantine stereotypes	gravity falls full episodes	

Comcast Confidential - Any unauthorized disclosure or use is strictly prohibited

Investigation of YouTube domain queries detected to remove Gibberish

We investigated **1,500** queries obtained from **Priority Set** (**Approach 2**: https://comcastcorp.sharepoint.com/sites/AppliedAI/_layouts/15/download.aspx?UniqueId=a1b09f165eee4caca2f42a3fe969eabd&e=DXeOXL) to check the quality of the set: how much of them are valid YouTube contents/search keywords and what are the examples of gibberish queries. Also, this investigation will enable us to find the patterns of gibberishes that can be detected by model.

Benefits

- Supports trending contents (ex. music, news, sports events) on YouTube better
- Supports YouTube channel name
  - (e.g.) ninja kids songs ninja kidz songs, sss sniper wolf sssniperwolf, j crew jkrew
- Handle various patterns describing the same target contents
  - (e.g.) **cocomelon**: coco mellon wheels on the bus, call camello, coco miller, cocomelon videos, cook amelon, cocomelon baby shark, coco malone, coco mellon baby shark, baby shark cocomelon, cocomelon bingo
  - (e.g.) **try not to laugh**: try not to laugh challenge, try not to laugh clean, try not to laugh tick-tock, not to laugh, turn up the laugh, turn on to laugh, clean try not to laugh, channel to laugh, why not to laugh, not to laugh challenge, try not to laugh videos, try not to laugh impossible, markiplier try not to laugh
- Suggest right contents for the query with ASR error
  - (e.g.) **elsa and anna**: elsia and anya, anya and elsia, elsia and anya, elsa and anya, elsia oneonta, onion elsia, lci narnia, on and elsa, amiens elsia, alsea and oneonta, elsia and oneonta, ~~lcn ony~~, oneonta elsia, elsia narnia, lcn anya, oneonta and elsia, elsie and anya

Issues

- Title of the contents or name of channels are sometimes very close to Gibberish pattern, which makes the distinction very hard.
- Some queries are words which are too generic to specify certain things, ideal to search all relevant contents on YouTube.
  - (e.g.) cakes, train, singing, graduation, etc.
  - (e.g.) name of the person (see below "ambiguous person names")
- Some queries are the words that should be reserved for command/conversation (see below)

Gibberish

music is, open, de son, how to make, sbl, vhc, add play, two songs, versus, abc's, frank, take, currently, walking, download, flip, a song, change to kids, the love, fleming, shoot, I want to go, yummy yum, adult me, i see, hi this, clean, for show me (probably fishy on me), say show (probably say so), what does, and beyond (probably trinity and beyond), true true, in the middle, z

(Note) Exclude the cases where "suggestion" exist and makes sense: "the score is" the skorys

Ambiguous Queries

sdtv, cyrus, mop, , infinite list, i am (channel/song exist), my story, else (singer)

ambiguous person names (still valid YouTube queries)

lexi, matt, mike, ali, bandy, asmal, tommy, thomas, corey, stewart, john, tom, antonio, benny, james, jeffrey

Queries reserved for Conversation/Salutation

hey (song exist), say, good night, okay, hi baby

Queries reserved for command

special, channels, volume down, the news, latest, content, my playlist, turn the channel, the tv, timers

Bad words

ass, fuck, suck, adult

Investigation of YouTube domain queries with many neutral (or mixed) user feedbacks

Motivation

Here are basic observations when we try combining signals of user experiences against YouTube domain queries.

- **YouTube queries with "Positive user experience"**
  - Most of these queries are related to the title of contents (or name of channel) which are available both in Xfinity and YouTube
  - We'll remove these queries from our YouTube predictions to avoid many overlaps in terms of suggestion (i.e. prefer TV to YouTube)
- **YouTube queries with "Negative user experience"**
  - Most of these queries are related to the contents (or name of channel) only available in YouTube.
  - We can improve the YouTube app experience by suggesting YouTube contents correctly **(Need to get rid of gibberishes)**
- **YouTube queries with "Neutral user experience"**
  - Lots of "Neutral" or "mixed" user experience means users are satisfied in some specific situations, while responses don't match well with user's intentions in the other situations.
  - These need to be investigated further to come up with better solutions

Investigation

(Case 1) Related to multiple contents

1. Max OR Maxxx (TV Series in 2020)
   a. Maxxx (TV Series in 2020)
   b. Max (2016 Movie with dog)
   c. Unimax (Channel)
   d. HBO Max (Channel)
2. Cat
   a. The cat (Movie in 1977)
   b. Genre like cat, cats
   c. Contents OR Genre: Cartoon cat, funny cat

(Case 2) Entity suggestions which is not satisfactory

1. Eminem (Rapper)
   a. Eminem entity page
   b. Songs by Eminem
   c. LMN (Lifetime movies)

(Case 3) Ambiguous query due to ASR error

1. Top
   a. Top (Movie from Prime Video)
   b. T.O.T.S. (Disney animation)
2. Call
   a. The call (Movie in 2013)
   b. College football
   c. call (agent), call home
   d. Caracol
   e. .....

(Case 4) Suggestion based on Genre is not satisfactory

1. Dance
   a. Contents related to "Dance"
   b. Sundance (Film Festival)
2. Cooking
   a. Contents related to "Cooking"
   b. Related channel: Food network, Cooking channel
   c. Related YouTube contents

(Case 5) Suggested contents needs payment

1. Grant
   a. Grant (TV Series i 2020)

Comcast Confidential - Any unauthorized disclosure or use is strictly prohibited

Refine rules for YouTube detection

From the investigation of current YouTube override set, we noticed some gibberish-like queries were included like these:

- music is, open, de son, how to make, sbl, vhc, add play, two songs, versus, abc's, frank, take, currently, walking, download, flip, a song, change to kids, the love, fleming, shoot, I want to go, yummy yum, adult me, i see, hi this, clean, for show me (probably fishy on me), say show (probably say so), what does, and beyond (probably trinity and beyond), true true, in the middle, z

New algorithm that utilizes tags as entity

We tried eliminating these gibberishes as many as possible by trying new rules which are more conservative in detecting YouTube contents. Followings are several strategies to improve the algorithm.

1. Utilize tag information in YouTube search result to eliminate short token queries (e.g. z) or incomplete queries (e.g. what does)
2. Special handling of queries with combination of keywords: tiktok renegade, elsa and anna toys

Here's the result with refined logic that eliminated 50% of gibberishes (17 / 34)

- music is, de son, vhc, add play, currently, flip, a song, change to kids, fleming, shoot, i want to go, yummy yum, i see, hi this, say show, what does, and beyond, z

Weakness of the algorithm

1. Popular contents without tag information were removed from valid set: (e.g.) baby shark cocomelon, bill maher, baking cakes
2. Contents with heavy ASR errors were removed from valid set: (e.g.) oneonta and elsia (Anna and Elsa)
3. Expression containing tokens in addition to keywords: (e.g.) babies dancing, A is for Adley (A for Adley)

Comcast Confidential - Any unauthorized disclosure or use is strictly prohibited

List of Orphan Domains

Domain	Description
Youtube	Youtube is the App-related domain representing YouTube contents, which can have one more of any of these domains list here (such content, music, news, etc)
Pandora	Pandora is the App-related domain representing music contents in Pandora, which can be modified as "music" domain in the future to represent all music contents.
Number	Number is the domain representing queries containing only digits which can be directly routed to TV agent. (Goal: Reduce traffic by not considering other agents)
News	News is the domain representing queries related to trending news
Gibberish	Gibberish is the domain that identifies nonsense queries that systems cannot understand, which can be categorized into these types: "Abbreviation", "Ambiguous", "Incomplete", "SR(Speech Recognition) Issue", "Repetition"
General Knowledge	General Knowledge is the domain representing queries by which user asks for information about people, events, etc. "Local search" is one fine-grained domain belongs to this.
Recommendation	Recommendation is the domain representing queries relevant to user request that leads to the suggestion of contents.
TV Guidance	TV Guidance is the domain representing question-type queries that request information about TV programs that Comcast provides.
Device Control	Device Control is the domain representing command-type queries necessary to control TV & Set-top Box.
Contents	Contents is the domain representing Xfinity Contents which TV agent can provide.
Apps	Apps is the domain representing command-type queries necessary to control Xfinity Apps.
Channel	Channel is the domain representing all channel queries.

In Production

Domain	Query Update Frequency
Youtube Detection of YouTube Domain queries	Daily via ALF
Number	Fixed set
Pandora	Daily via ALF
Gibberish	Fixed set
News	Daily via ALF
Local Search	Fixed Set
Device Control	Daily via ALF
Xfinity Mobile	??

Production Ready

- ...

Under Research

- News
- General Knowledge: Clustering Approach (II) to identify General Knowledge domain
- Recommendation
- Apps
- TV Guidance

- Contents
- Device control
- Channel

Real query investigations for unhandled domains and gibberish

Gibberish

Incomplete

Set TV for, I need to get, I want to start, For a, How you, I don’t know to, Do you account, What you channel, I negotiate and Utah, Any good movies on, Are you can you can, Hey Google find the movie that I, Movie ready for, Why is the world of, Are getting by ("Getting by": American sitcom), Go to, To be

Repetition

Show me TBS Hey Xfinity broadcast, Please Netflix Netflix please, Netflix I mean my exit exit exit, Yes dear on YouTube one fish two fish, YouTube Spotlight and Friends are being mean to you you still don't care, Move Smart Remote Smart Remote resume

Irrelevant (Non-sense)

Please find what I’m writing, Can’t stop this feeling, I want machine menu, I’m a sucker for you, You're a hoe, Please stop the described video, Yeah I am Star, Turn to Bluetooth device, I'm a evil little Bratz me, Dad why do you can you so weird, How to go for my little bit of hair from my lots of Hair by using baby dummy Bob, Show me not play, It is the power of the Jets on yelp help, I found a fiancé USB drive, Know what's the name, Cancel Single seven, Bring up future lottery numbers, Rentable apartment at Disney, Go to repeating to dog, Heroes is the rock.com, What's on the Big, Are you Game? My best friend my best holiday, Play the card please, Fix my Alexa, I guess again this, Come come my remote, What is a lie, When Santa Thursday to Friday and Sunday Monday episode and then decided, We sure had to sit three freaking days to listen to the Democrats lies over and over the same thing, 8 PM, Didn't even know you put yourself tomorrow show you gotta keep, Crisis on Hey Xfinity are, You're right this is that that's part of it that is it, There will say listening, What's up what's up the website and

Not meaningful with repetition

Pupu Pupu Pupu Pupu, Dude Perfect rocketship blah blah blah blah blah stop, No no no no, Tankian Hey Xfinity poop, Alisi Shin-Wook Ebenezer Lala Lala Lala Lala, Up up down down left right left right, Long Long Long Long, The right the right of my no no no exit stop

Ambiguous (Unknown tokens)

X FX, Hear hi, PD please, P VS, DJ Park, YC, GI, CL, Penguin penguin upcoming, B 24, 13 W., D a R, TG, A H2, 06 CM, KC KC, DS, L

New domain

Music

50s music on, i heart radio station, Give me baby shark when you, Pet Shop Boys One More Chance, ID Thunder, Mellow song, Play the song standing on the verge of Getting On by the funkadelic's, Play Thompson a girl, I love it when you call me Senior Rita (señorita), Lady Gaga I'm alone in my house, 80s country love song, Play Perry Como's Greatest Hits, Wasted days and wasted nights song, Thompson Twins Take on Me, Stephen Bruton on it is what it is, I wanna What Love Is foreigner, Bob Marley Exodus song, Music by Queen (==> The Music: Getaway), Listen to music (==> The Doobie Brothers: Listen To The Music)

Youtube Search

Qué tiemble de Daddy Yankee, Tick-tock videos YouTube pretty ladies, Funny Videos on Peppa Pig, Tom and Jerry is you is or is you ain't my baby, Jailbreak new vs. pro, Number blocks, I'm still gonna send it, Eric Clapton (==> Eric Clapton: Life in 12 Bars), Three danger (==> BROWSE genre: danger), Play again (==> "Bare Naked Survivor: Again", play <movie>), Sportsnet Central (==> No result found), Ring (==> The Ring, Probably "Ring" by Selena Gomez"), Qué hora es? (==> Time, Can be "Que Hora Es? Part 1), What's your name? (==> What's Your Name: In the Style of Lynyrd Skynyrd, more results on YouTube), African dance (==> Joffrey: Mavericks of American Dance)

Other Contents (Provider)

I don’t think there Disney, Audio English, Amazon music Atlantis Morissette, Amazon prime Network, Renegade music on Amazon, Hi Siri turn the volume back on please

Question: General Knowledge

Is New Years over in France, What’s flamingos, Why is Johnny a communist, How old is Alex Trebek, What does BTS stand for, What are the woodpecker

Question: Guidance

Shows on TLC looking for biological family, What time is the show over, Where is Chrisley (Chrisley knows best), How many episodes of the old are there, What network is airing the Kobe Bryant special, What time do the global awards come on and what channel, What channel is American music awards on, What time is the awards on, What movies are on at 8 o'clock, Is the Houston Rockets playing right now, When does Forrest Gump playing on the star channel, What channel is the football game on, When does the Super Bowl start, What's on TV tonight, News that's on now, NBA on now

Question: Sports

Who scored the most points in NBA, Who won the NFC AFC game, When is Rafael Nadal play his game Tennessee, How did Tiger Woods do at the State Farm insurance, How many points Zion Williamson have, Who is playing on the pelican today,

Question: Math

What's 9+3

Recommendation

Any good movies on, Movies like shazam, Sing-along songs on Nickelodeon, Free kid movies half an hour, Movies for Family Night, Movies starting with the letter J, Show me something I might like to watch, Really fun and exciting movies that are free for kids, Show me movies on The Big Valley, Movies that are about law firms and quizzes, Free movies with no ads, Free Scooby Doo cartoon

Device Control (Timer, Recording, Auxiliary devices, etc)

audio setting(s), Show timer, Make the picture brighter, Set timer for 830, Set timer for 15 minutes, Watch it tonight, Record game on ESPN, Record AMC awards, Take A Picture of me and I'm in front of the TV that is this TV, What time is it Question Mark, English please, Speak English, Change language, fox espanol, HD fox, Play in 4K, Record the Windsors, How to schedule

Visceral

What am I pay for the shit service, Fuck you fuck my remote mother fucker fuck you a smelly, Never right with this thing anymore

Travel

Where are the Cayman Islands, I wanna Ride

News (Trending)

California status, Show me news Kobe Bryant, Update on Coby Brown, Up-to-date news on Kobe Bryant's death, The Iowa caucus, Corona virus

Sports (Trending)

UConn women's game, USA vs. Costa Rica, US Women's Soccer team, Super Bowl Chiefs vs. 49ers

Lottery

Powerball, Power ball

Food

Skewers best place ever, O défi

Others (Need improvement)

Number/Spacing/Special token

Minions two, Trolls two, Sing two, Avatar two, OJ 25, Trolls The Beat goes on (==> Trolls: The Beat goes on), CW 50 (==> CW50), Watch What Happens: Live (==> Watch What Happens Live)

Spell error

what is my wife?, Are Law & Order, Baby game vs. Kansas (==> Navy game vs. Kansas), Force Gump, Leiden's mystery Playdate (==> Ryan's mytery playdate), CANA L St., CANH L St., Mama mental (Probably "Mapa mental"), Show me the Kingsmen Secret Service, Strawberry Shortcake and Netflix (and ==> on), L a Lakers, All the CD you two (==> U2), KS PS (==> KSPS), MNS BC MNS BC, I will caucus (==> Iowa Caucus), Cartoon net (==> Cartoon Network), The Habs and have not (==> the haves and have nots)

Repetition

Bowl bowl game, What is what time is it,

Guardword

netflix seven days to utopia, Netflix Ryan Reynolds, Free Scooby Doo cartoon, Free secret life of pets two

Other Languages

Hallmark hoy fue, Amar a muerte (TV Series), Viendo

Others

Show geographics,

New directions for OpenNLP

1. New Fine-Grained Domains (Independent from agents)

Limited transition (for N/A queries)

Fix N/A errors by doing semi-automatic annotation (Pattern discovery)

- Problem: What agent should handle this? Agent-dispatcher?

Full transition

Build annotation data semi-automatically by applying clustering to all queries for new fine-grained domains and fixing errors manually

2. Reinforcement Learning with XRE data

Improve the performance of OpenNLP by utilizing XRE data as reward signals

- Problem 1: Part of error is attributable to Corti-agents, not to OpenNLP
- Problem 2: Reinforcement learning requires real production systems setting for learning/evaluation

3. Reduce F/P of Domain Selection using Context

Reduce the traffic of OpenNLP by utilizing "App Context" input

- Problem: Annotation is hard as we need to identify different domains/answers for the same query in all different contexts. Need another signal that indicates whether the selected domains/answers are satisfactory or not.
- Approach
  - [x] Collect queries that return multiple domains
  - [x] Get the distribution of contexts for each query
  - [x] Investigate how context information can help in deciding the final answers for collected queries

4. Smart Reply with Synonyms

Build the list of synonyms (different ways of representing contents, queries with minor ASR error) for smart reply

- Problem: What's the boundary of OpenNLP?

Comcast Confidential - Any unauthorized disclosure or use is strictly prohibited

Using Context to inform OpenNLP Research

JIRA Epic is VREX-9285

Annotated Data Produced

https://comcastcorp.sharepoint.com/:x:/r/sites/AppliedAI/Shared%20Documents/Projects/Corti/domains/opennlp/annotation/queries_for_annotation-JAN01.xlsx?d=wf9a10948e0ba47c8aa390303e14f3544&csf=1&e=bYSAEc

Using Databricks to get the Context from VREX

If I am new to the team, how do I start using Databricks?

To get access to Databricks you need to fill out the form here and submit the request for access.

https://tpx.sys.comcast.net/servicedesk/customer/portal/45/create/915 (Summary example: "Add Scott Roam to Applied AI / AIQ, Requirements: Access request <email address>, Manager: Jan Neumann)

Once approved, you can access your Databricks account at http://dx-comcast.cloud.databricks.com/

You can create you notebook by specifying `language` and `cluster` and then start coding.

How can I collect the Context from VREX using Databricks?

Following s3 bucket contains the Vrex logs (with context) for specific date and hour.

- path = 's3a://atlantic-production-cloudbridge/data/deap/raw.mirrored.vrex.VoiceCommand/' + <date_str> + '/'+ <hour_str> +'/*/*/*'

You can simply use SparkContext to get the file from the bucket like following.

- vrex_logs_raw = sc.textFile(path)

Here's the context processor code you can check. (**Need to get permission**)

- https://dx-comcast.cloud.databricks.com/#notebook/6913863/

Examples of Context (From data from DEC 6 - 9 2019)

Context only contains one attribute "appFocuses" at this point. For further information, you can see the following page. Context Analysis

Collect data for training (annotation)

One of the challenges in the research (M/L domain selector that use "Context") is the non-existence of labeled data. We can assume that the final response (i.e. output of answer selector) of current production systems is the true label for domain selector to mitigate the issue. However, we cannot do correct evaluation with somewhat biased, inaccurate training data collected from production systems. Therefore, we need additional annotation work to collect labeled data, part of which will be added to training data while others will be used as evaluation data. Annotation work will be done by using Excel spreadsheet for this research in following format.

QUERY	COUNT	CONTEXT	PREVIOUS QUERIES	LANGUAGE	PARTNER	DOMAINS
NBC	12289	amazonPrime	Amazon Prime	eng-USA	comcast	tv
A&E	4526	xre:event:STA_054:entity_page	Discovery;Secrets	eng-USA	comcast	tv
.....		.....		.....	.....	tv;insights

Previous queries among all the columns above can be obtained in following way:

- Group queries by device_id first to collect all queries from same device
- Group queries into session by checking time interval (currently, 45 seconds)
- Get all previous queries in the same session (for the target query)

In order to collect data more effectively, we focus on collecting following samples:

1. ~~which are most frequent, i.e. has high priority (automatic: no requirement for manual annotation)~~
2. which belong to one domain clearly by domain selector returning only one domain and answer selector accepting it (automatic: no requirement for manual annotation)
3. which belong to one domain clearly while domain selector returns multiple domains (automatic: no requirement for manual annotation)
4. which are ambiguous, technically speaking, located around the boundary of two domain spaces (need annotation)
5. which are identified as "fail" by sessionized queries and XRE signal (Will be done in the next stage)

There are 3 different situations which belong to "ambiguous" category like following.

- Domain selector decides multiple domains and more than one agent produces "SUCCESS" response
- Domain selector decides one or multiple domains and all selected agents produce "None" response
- Domain selector decides nothing

RESEARCH

What is the research question we are addressing?

The domains selected by the domain selector are sometimes misleading, especially with short queries that do not contain much context inside. The goal of this research is to investigate whether or not additional context information is useful in making decisions of domains. The model which takes input of previous utterances (obtained from session queries) will be also tested.

Effects

The selection of right domain increases recall of selected domains, which will eventually lead to correct answer to the query. Also, further clarification of domain from context will reduce the number of domain candidates from domain selector (i.e. increase in terms of precision) and consequently reduce the executions of Corti-agents and save operation cost.

How are we addressing it?

While current domain selector only process text portion of given query, we can extend the input features for the model to be more accurate. The first possible input is context data, i.e. context in which a user is at the time of speaking. The second part is the history of user utterances (or previous utterance) from which we can assume the context indirectly. In short, the likelihood of target domain will be conditioned on the current context like P(yt | ut, context) or previous domain and utterances like P(yt | yt-1, yt-2, ... ut, ut-1, ut-2, ...) in addition to current query. We can use statistical model or M/L model (i.e. recurrent neural networks) that allows us to consider additional inputs to solve problems.

Approach

1. Build codes to collect/persist target data we want (context information, a series of utterances in a session)
2. Investigate target data and get insights
3. Collect annotated data (Refer to the contents above: "Collect data for training")
4. Build new model that ingests target data
5. Train current model and new model
6. Build metric and perform evaluation for current model and new model

Challenges

1. There's no data containing true labels (i.e. domain). We have to assume that the responses of current production systems are true. We need to collect annotated data at least for the purpose of precise validation.
2. ~~The model will be more complicated if different type of context data exist~~. Only one type of context data exist at this point after data investigation.
3. How we will decide one session will affect the behaviors of systems.
4. As current domain selector requires low latency, using data such as previous utterances will make the model more cumbersome.

What have we learned?

Context Analysis

Distribution of contexts (with duplicates)

Context	Frequency	Ratio (%)
app:MEDIATUNE_500	1199894	69.934978
app:xre:event:STA_054:entity_page	389507	22.702142
app:YouTube	56082	3.268700
app:NetflixApp	49172	2.865955
app:amazonPrime	4459	0.259890
app:KIDSMODE	4049	0.235993
app:SportsApp	2085	0.121523
app:LowerThirds-Sports	1788	0.104212
app:XRE_GUIDE	1737	0.101240
app:Tubi	1415	0.082472
app:PandoraApp	1214	0.070757
app:LowerThirds-Weather	735	0.042839
app:YouTubeKids	603	0.035145
app:siX1App	531	0.030949
app:AmazonMusic	397	0.023139
app:iHeartRadio	346	0.020166
app:myaccount	198	0.011540
app:MusicChoice	195	0.011365
app:TransgamingPortal	160	0.009325
app:sports12c	134	0.007810
app:xfinityhome	101	0.005887
app:pluto	96	0.005595
app:dazn	88	0.005129
app:holiday2019	78	0.004546
app:AppStore	74	0.004313
app:forthethrone	62	0.003614
app:HattrickApp	57	0.003322
app:NPROne	49	0.002856
app:ChannelStoreApp	48	0.002798
app:xumo	37	0.002157
app:troubleshooterApp	36	0.002098
app:PictureFrame	33	0.001923
app:djchristmas	25	0.001457
app:XITE	21	0.001224
app:LowerThirds-xFi	19	0.001107
app:shortFormVideo	16	0.000933

Comcast Confidential - Any unauthorized disclosure or use is strictly prohibited

app:LowerThirds-Stocks	15	0.000874
app:TransgamingPortal_CrossyRoad	14	0.000816
app:SlingTV	12	0.000699
app:voicemail	12	0.000699
app:roadfury	11	0.000641
app:xre:event:CTX_1000:menu	11	0.000641
app:UniversalKids	9	0.000525
app:LowerThirds-Song	8	0.000466
app:HappyKids	8	0.000466
app:DreamworksGames	7	0.000408
app:tombrunner	6	0.000350
app:footchinko	6	0.000350
app:LowerThirds-Horoscopes	6	0.000350
app:xre:event:CTX_1000:entityInfo	5	0.000291
app:poll	5	0.000291
app:Fandango	5	0.000291
app:repackagerx1	5	0.000291
app:FilmRise	4	0.000233
app:NBCSportsGold	4	0.000233
app:xre:event:CTX_1000:browse	4	0.000233
app:TransgamingPortal_WorldPokerTour	4	0.000233
app:Sportsnet	3	0.000175
app:yetigetaway	3	0.000175
app:tetris	3	0.000175
app:Fawesome	3	0.000175
app:zoneify	3	0.000175
app:hallmarkholiday	2	0.000117
app:Hulu	2	0.000117
app:myacct-outage	2	0.000117
app:kshalloween	2	0.000117
app:halloweengameroom	2	0.000117
app:Zoneify	1	0.000058

Distribution of contexts (without duplicates)

Context	Frequency	Ratio
app:MEDIATUNE_500	137961	55.571623
app:xre:event:STA_054:entity_page	73513	29.611533
app:YouTube	27418	11.044156
app:NetflixApp	5458	2.198519
app:amazonPrime	838	0.337552

app:KIDSMODE	520	0.209460
app:PandoraApp	459	0.184888
app:Tubi	355	0.142996
app:SportsApp	332	0.133732
app:LowerThirds-Sports	313	0.126079
app:LowerThirds-Weather	231	0.093048
app:XRE_GUIDE	184	0.074116
app:YouTubeKids	128	0.051559
app:iHeartRadio	94	0.037864
app:myaccount	54	0.021752
app:siX1App	49	0.019738
app:MusicChoice	48	0.019335
app:AmazonMusic	38	0.015307
app:TransgamingPortal	36	0.014501
app:pluto	28	0.011279
app:sports12c	28	0.011279
app:xfinityhome	23	0.009265
app:HattrickApp	17	0.006848
app:holiday2019	17	0.006848
app:ChannelStoreApp	11	0.004431
app:troubleshooterApp	10	0.004028
app:AppStore	9	0.003625
app:dazn	9	0.003625
app:djchristmas	7	0.002820
app:xumo	6	0.002417
app:PictureFrame	5	0.002014
app:LowerThirds-Stocks	4	0.001611
app:NPROne	4	0.001611
app:XITE	4	0.001611
app:TransgamingPortal_CrossyRoad	4	0.001611
app:forthethrone	3	0.001208
app:tombrunner	3	0.001208
app:FilmRise	3	0.001208
app:voicemail	3	0.001208
app:xre:event:CTX_1000:menu	3	0.001208
app:roadfury	2	0.000806
app:hallmarkholiday	2	0.000806
app:LowerThirds-xFi	2	0.000806
app:Hulu	2	0.000806
app:shortFormVideo	2	0.000806
app:xre:event:CTX_1000:browse	2	0.000806

app:Fandango	2	0.000806
app:HappyKids	2	0.000806
app:NBCSportsGold	2	0.000806
app:xre:event:CTX_1000:entityInfo	2	0.000806
app:Fawesome	2	0.000806
app:UniversalKids	1	0.000403
app:yetigetaway	1	0.000403
app:LowerThirds-Song	1	0.000403
app:TransgamingPortal_WorldPokerTour	1	0.000403
app:SlingTV	1	0.000403
app:DreamworksGames	1	0.000403

Comcast Confidential - Any unauthorized disclosure or use is strictly prohibited