Data Curation, Crowdsourcing, Integrations

2/12/2020
Objectives

- Create effective crowdsource jobs to gather data.
- Learn about basic data curation needs
- Understand how to integrate with Clinc
Crowdsourcing
Gathering Classification Data

Scenario

- Set up a scene
- Ask worker what would they say in this scenario
- Gathers a wider data set

Scenario-driven
Suppose you have an intelligent assistant that can handle ______. Given the following scenario, supply 5 creative ways of asking the intelligent assistant to help with your situation.

“______________________________”

Paraphrase

- Provide examples
- Ask worker to provide different expression with similar meanings
- Gathers targeted data

Paraphrasing
Given the following sentence, supply 5 creative ways for rephrasing it and expressing the same meaning.

“______________________________”
Classification Job Examples

Scenario

Scenario-driven
Suppose you have an intelligent assistant that can handle ______. Given the following scenario, supply 5 creative ways of asking the intelligent assistant to help with your situation.

“___________________________ JOB PROMPT ________________”

Generic

“You’re looking for a specific set of transactions in your bank account”

“You want to learn about your spending history”

Specific

“Tell me about my transactions during the month of November”.

“What was my spending on food and drink?.”

Paraphrase

Paraphrasing
Given the following sentence, supply 5 creative ways for rephrasing it and expressing the same meaning.

“___________________________ JOB PROMPT ________________”
Crowdsourcing Classification

Tips

● Gather small data sets first
● Provide multiple examples
● Use diverse, creative examples
● Use different slots
● Track your jobs and iterate
Classification Data Curation

- **Example:** “Pretend you need to transfer money from checking to savings, what would you say”
  - Intent: transfer
  - Slots: to_account, from_account, amount

- **Common crowdsourcing problems:**
  - **Noisy data:** data that is not relevant
    - “qwerty”
    - “Online bank transfers. Log in to your online account and select the option for making a payment.”
  - **Incorrect data:** wrong intent, or not in-scope
    - “Can I speak to someone in person?”
    - “What is my balance?”
  - **Homogeneous data:** data that only contains relatively few unique words
    - “Transfer from checking to savings”
    - “Transfer money from checking to savings”
    - “Transfer money from savings to checking”
Fixing Homogeneous Data

- **Homogeneous data**: data that only contains relatively few unique words
  - “Transfer from checking to savings”
  - “Transfer money from checking to savings”
  - “Transfer money from savings to checking”

- Classifier will be “overfit” to certain words (bad)
  - Overfit: only queries with certain words will work

- This can be alleviated by:
  - **Using varying crowdsourcing prompts**
    - “Pretend you need to transfer money from checking to savings, what would you say?”
    - “Pretend you want to move money between two of your bank accounts, what would you say?”
  - **Using slot replacement**
    - checking → [checking, savings, money market, college savings, Premier account, …]
  - **Using synonym replacement**
    - transfer → [transfer, move, ship, relocate, shift, siphon, …]
Curating Classification Data

**General Concepts**

- Ensure that all utterances are within competency scope
- **Do not** delete typos, slang, incorrect grammar, etc.
- Delete copy cat utterances
  - “Is this teaching the AI something new?”
- Consider the AI’s response --- does the training data cover all avenues to AI’s response?
Classification
Uniqueness Sorting Tool helps you identify irrelevant utterances in classification data.

For most unique utterances:
- Are there irrelevant data?
- Are there underrepresented data (i.e. specific keywords) we need more of?

For least unique:
- Are there phrases being overfitted?
The uniqueness sort tool works by first mapping each query to a high dimensional semantic embedding space. A mean embedding representation is then computed, and finally the tool sorts each query by distance from the mean embedding.

A query’s uniqueness score is the query’s distance to the mean embedding.

Uniqueness sort draws attention to the most unique utterances so that you can identify weak spots in data or potentially mislabeled or out of scope data.
Uniqueness Sort Tool

- Uniqueness Sort (or anomaly detection) is a data insight tool that Clinc provides to analyze classifier data to identify errors or underrepresented training samples
  - Errors are training samples that are either mislabeled or unnecessary noise
  - Underrepresented training samples are good but unique compared to the rest of the samples
Accessing Uniqueness Tool

- Navigate to the intent classifier page
- Select Sort By -> Uniqueness
Data curation is an iterative process!
- Gather data
- Curate
- Test

Use fine-grained intent labels

Test your classification first! Similar to incremental development, you want to test what you build in as granular chunks as possible. You want to test to make sure that your classifications work as expected before doing too much more additional work to minimize potential re-work.
Front End Integrations
What is a Clinc Integration?

Enabling users to experience Clinc’s rich conversational experiences and sophisticated NLU across various channels and devices.
How do Apps Integrate with Clinc?

Configuration
- Configuration data for the app itself and Clinc’s APIs

Authentication
- Methods and settings for linking external accounts and authenticating requests to Clinc’s APIs

Query Endpoint Connection
- Sending and receiving HTTP requests to Clinc’s APIs that use the AI version and Business logic
App Configuration

Clinc Configuration
- UUID of the AI Version being used
- Credentials to request auth tokens
  - Username
  - Password

Integration Specific Configuration
- Application set up
  - Example: Creating an Alexa Skill or Web app
  - Any configuration needed by the application itself.
App Authentication

Clinc Authentication
- Clinc’s APIs require Bearer Tokens be set in the request header
- To obtain a token, HTTP POST to `/v1/oauth` with credentials
- Credentials set in body

External Authentication
- External user accounts can be linked in the integration by passing auth tokens in the body to the query endpoint

HTTP POST /v1/oauth

Header

```
Content-Type: application/x-www-form-urlencoded
```

Body

```
username : bob
password : $3kr3tp@$w0rd
grant_type : password
```

Response

```
{
   "access_token": "acdefghijklmnopqrstuvwxyz12345",
   "expires_in": 36000,
   "token_type": "Bearer",
   "scope": "user_read user_write query",
   "refresh_token": "12345678901234567890123456789"
}
```
Query Endpoint Requests

To use the Clinc API, HTTP POST to `/v1/query`

Set Bearer token in the header

**Key items in request body:**
- `query`: User's text query for processing by the AI and Business Logic (BL)
- `ai_version`: UUID of the AI version being used
- `dialog`: Optional token used to maintain conversation state
- `token`: Optional token used to link external accounts - can be used in BL

**HTTP POST** `/v1/query`

**Header**
- `Content-Type: application/json`
- `Authorization: Bearer abcd1234abcd1234abcd1234abcd1234`

**Body**
- `query`: "Can I transfer $100 from savings to checking?"
- `ai_version`: "123abc123abc123abc123abc123abc123abc"
- `lat`: 42.2730207
- `lon`: -83.7517747
- `time_offset`: 300
- `dialog`: "abcdefg123456789"
- `device`: "alexa"
slots
● Object of keys and values collected by the conversation
● Useful when application needs neat parseable data that has been gathered from AI Version and Business Logic.

dialog
● Token to maintain conversation state
● Send dialog token in the body of request to the Query Endpoint to preserve data in the conversation so far.
A state is a node on the state graph (seen in the platform) which can be classified into given an intent.

Useful when application needs information on what the user is doing.

Example:
- User opens the application and says “Hello”
- `classifier_state` is `greeting`
- The application could display a “Welcome” page or animation.
Query Endpoint Response Continued

**speakableResponse**
- Text response suited for a voice experience
- Value is passed to Text to Speech (TTS) services

**formattedResponse**
- Text response for displaying and reading in the application
- Useful when the response is too long and bulky to be heard aloud or contains sensitive data that might not want to be broadcasted
Other Considerations & Example Project

Automatic Speech Recognition (ASR) and Text to Speech (TTS)

- Some integrations, like web app integrations require external libraries and APIs for TTS and ASR for a voice first experience.
- WebKit Speech, Google Cloud Speech, Responsive Voice

Example Project

- Alice wants to build a Clinc Integrated Web Application
  - ASR
  - TTS
  - Authentication
  - Query Endpoint
Utterance
I would like to check the balance in my checking account.

Response
As of January 3rd 2020 you had a total of $16,981.84 in your Checking accounts.