Dialogue corpora

NPFL070

December 11, 2019
1. Intro

2. Task oriented

3. Chit-chat

4. QA
What is dialogue

Sample conversation

Hello, how may I help you?
I am looking for a cheap restaurant in the city centre.
There are over twenty cheap restaurants. Which cuisine do you prefer?
I like chinese food.
Golden palace is a cheap restaurant with good ratings.
That sounds good, can I have an address and phone number please?
...

...
Dialogue tasks

- What is the use case?
  - task-oriented dialogues
  - ”chit-chat”
  - Question Answering (QA)

- Subtasks
  - Natural Language Understanding (NLU)
  - Dialogue State Tracking
  - Dialogue Policy
  - Knowledge Base information retrieval
  - Natural Language Generation (NLG)
  - (ASR, TTS)
[credit: A Survey of Available Corpora for Building Data-Driven Dialogue Systems by Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, Joelle Pineau]
Terminology

- **turn** - one usr/system utterance
- **slot** - unit of semantic information, type=value
- **intent** - desired user goal
- **action** - system action

Example

I am looking for cheap chinese food.

\texttt{inform(pricerange=cheap), inform(food=chinese)}
Evaluation

- **Intrinsic**
  - NLU, State tracking - classification, i.e. accuracy, precision, recall
  - Dialogue success - were all the requests fulfilled
  - entity match rate - were relevant information provided?
  - BLEU - NLG, end-to-end setups

- **Extrinsic**
  - Human rating - experts, crowd platforms (can be problematic)
Dialogue dataset types

- **Modality**: written, spoken, multimodal
- **Collection process**:
  - human-human
    - real/scripted
  - human-machine
  - automatic (machine-machine)
- **Domain**
  - limited (closed) vs. open domain
Specific problems of dialogue data resources

- the central problem: unlike vast majority of NLP tasks, dialogue management is hard to decompose into independent subtasks, as each turn in a real dialogue is extremely sensitive to the previous turn(s)

- as a consequence, a man-machine dialogue typically quickly diverges from an authentic dialogue

- the fact that a dialogue composes of a sequence of turns, each of them corresponding to a few natural language sentences (i.e., the branching factor is astronomic), implies a HUGE search space ... 

- ... which is impossible to cover sufficiently by any authentic training data

- (some other NLP tasks such as machine translation also face huge search space, but dialogues are worse because of the sequential nature)
Collection process

Expert collection

- Good acoustic conditions, high level of control
- Usually very costly, high quality
- Scripted or Wizard-of-Oz scheme
  - Participants still talk to the system (machine).
  - The system is secretly controlled by another human.
  - Desired because people behave differently when talking to machine
Collection process

- **Web crawling**
  - fast, cheap
  - difficult to organize
  - prone to errors
  - often not real dialogues (tweets and replies etc.)

- **Crowdsourcing**
  - untrained workers employed through some kind of data collection platform
  - Crowdflower, Amazon Mechanical Turk
  - compromise in terms of cost and quality
Data labels

- One typically needs some data labelling (for language understanding, policy decisions).
- audio transcriptions
- semantic annotation (intents), (named) entity labelling
- other: POS, hypotheses
- experts, crowdsourcing, semi-automatic

Example

I want to fly from New York to San Francisco on Friday morning. request(from=NY, to=SF, date=Friday, time=morning)

There are two airports in NYC, JFK and LaGuardia. Which one of them do you want to depart from?

actions={ask_airport(), inform_multiple(JFK, LGA)}
Task (goal) oriented systems have defined goals that should be accomplished (book a restaurant, find a flight connection, find a sightseeing place).

The system’s task is to ask for the restrictions and user preferences and provide options.

Usually there is a domain-specific ontology, i.e. a priori knowledge.

Chit-chat systems however don’t need to accomplish anything.

The purpose is to mimic human behavior or keep the user entertained.

Both can use knowledge bases, i.e. database of facts.

There can be some overlap.
DSTC 2 (3) (2013)

- Dialogue State Tracking Challenge
- State = set of current slot values, possibly additional features
- human-computer, restaurant reservation system
- 3000+ dialogues
- DSTC 2 (2013) considered a benchmark for a long time
- Apart from state also turn-level annotations; language understanding = recognized slot values + intent
- included ASR hypotheses
- http://camdial.org/mh521/dstc/
MultiWOZ (2018)

- multi-domain, 10k+ dialogues in total
- state and actions annotations
- human-human; Wizard-of-Oz scheme
- database included
DSTC 1, Let’s go

- Let’s go - over 170k dialogues, transcribed
- DSTC1 subset of the corpus, state annotations
- public transport domain
- https://github.com/DialRC/LetsGoDataset
1936 conversations collected in Wizard-of-Oz fashion
Complex dialogues about flight and hotel reservations
Frame tracking - generalized state tracking, considering more constraint values in parallel
https://datasets.maluuba.com/Frames
- 3031 dialogues in 3 domains
- car assistant and driver
- human-human interaction
ATIS, DSTC6+

- Air Travel information services
- Human-machine, 774 conversations
- Dialogue State–Tracking Systems Technology challenge
- 2017 DSTC 6, 2018 DSTC 7, …
- Each year set of tracks & new dataset
Chit-chat: spoken corpora

- Collected dialogues on various topics, usable also for speech recognition
- Switchboard (1992) - 300h, telephone speech
  http://groups.inf.ed.ac.uk/switchboard/
- British National Corpus (1992) - 1000h, various sources
  http://www.natcorp.ox.ac.uk/
- Ami Corpus (1997) - 100h, meeting records, good quality
  http://groups.inf.ed.ac.uk/ami/download/
Chit-chat: written corpora

- Twitter customer support corpus
  - over 3 million tweets & replies
  - https://www.kaggle.com/thoughtvector/customer-support-on-twitter

- Ubuntu dialogue corpus
  - 930k dialogues
  - humans chatting about technical problems with Ubuntu operating system
  - https://github.com/rkadlec/ubuntu-ranking-dataset-creator
**Chit-chat: written corpora**

- Reddit all comments
  - 1.7 billion comments on Reddit discussions
  - [https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/](https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/)

- Movie dialog Dataset
  - 3 million short dialogues on movie recommendations
  - part of the bAbI project
  - [https://research.fb.com/downloads/babi/](https://research.fb.com/downloads/babi/)

- OpenSubtitles
  - human-human scripted dialogues
Cambridge RNNLG
- restaurants, hotels, laptop, TVs
- crowdsourced

E2E NLG data
- restaurants (bigger)
- more complex
- partially based on images
Question Answering

- knowledge retrieval
- text understanding, reasoning
- The ”dialogue” (conversation) aspect is not as important as providing the relevant facts and proving understanding.
Facebook bAbI project
https://research.fb.com/downloads/babi/

Sample

context: John gave a ball to Stephen. Stephen went to kitchen.
Q: Where is the ball?
Question answering

- WikiQA
- Yahoo QA