Data Augmentation for Conversational AI

The Web Conference 2024

Tutorial website

Presenters

Heydar Soudani
PhD Candidate
Radboud University
heydar.soudani@ru.nl

Evangelos Kanoulas
Full Professor
University of Amsterdam
e.kanoulas@uva.nl

Roxana Petcu
PhD Candidate
University of Amsterdam
r.m.petcu@uva.nl

Faegheh Hasibi
Assistant Professor
Radboud University
f.hasibi@cs.ru.nl

Supplementary Material

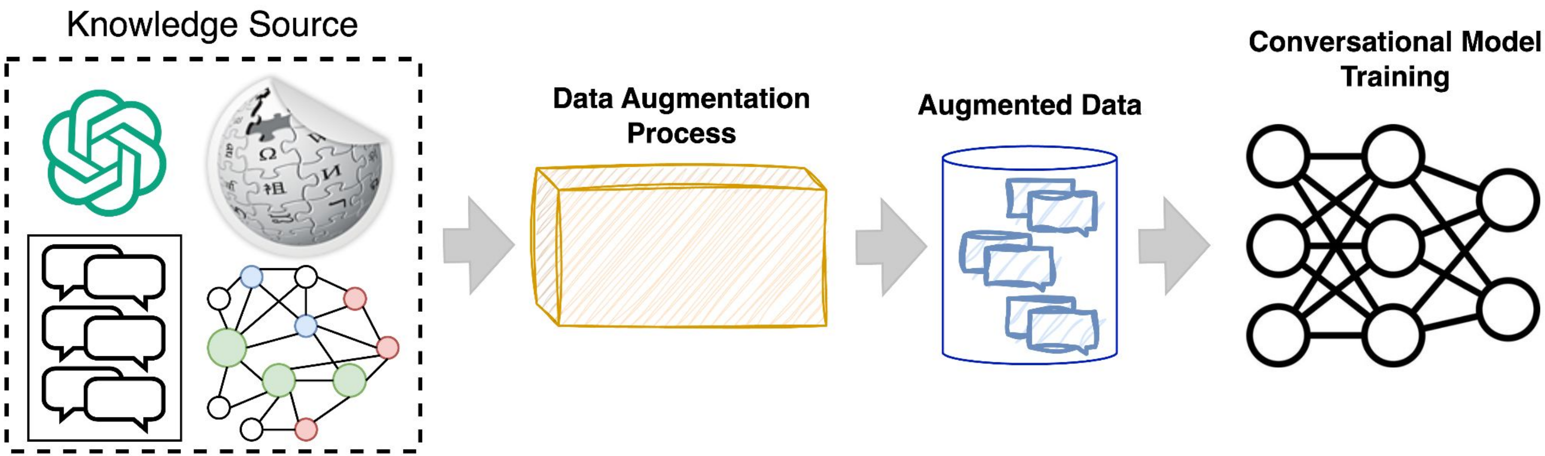
Website: https://dataug-convai.github.io/

A Survey on Recent Advances in Conversational Data Generation

HEYDAR SOUDANI, Radboud University, The Netherlands
ROXANA PETCU, University of Amsterdam, The Netherlands
EVANGELOS KANOULAS, University of Amsterdam, The Netherlands
FAEGHEH HASIBI, Radboud University, The Netherlands

Recent advancements in conversational systems have significantly enhanced human-machine interactions across various domains. However, training these systems is challenging due to the scarcity of specialized dialogue data. Traditionally, conversational datasets were created through crowdsourcing, but this method has proven costly, limited in scale, and labor-intensive. As a solution, the development of synthetic dialogue data has emerged, utilizing techniques to augment existing datasets or convert textual resources into conversational formats, providing a more efficient and scalable approach to dataset creation. In this survey, we offer a systematic and comprehensive review of multi-turn conversational data generation, focusing on three types of dialogue systems: open domain,

What is This Tutorial About?



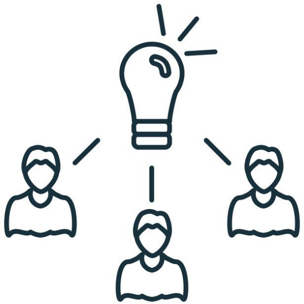
The process of artificially expanding the size and/or the diversity of training data for conversational AI models.

Why "Data Creation" **for** Conversational **AI?**

Because of the Data Scarcity issue

Data Scarcity in Conversational AI

- **Large neural models** are heavily used for dialogue systems
- Access to **large amount of training data** is key to the success of these models
- **Crowdsourcing** is a common approach to create large scale datasets



Crowdsourcing Limitations

- Expensive
  - CoQA spent 3.6 USD per passage on crowdsourcing
  - Total cost: ~ 30K for ~8400 passage and 127K QA pairs
- Time consuming, hard to scale up (Wu et. al. 2022)
- Generating data for conversational tasks is often complex (Gu et al., 2021)
  - High cognitive load for workers, low quality data
  - E.g., Preference Elicitation dialogues (Radlinski et al. 2019)

Data Augmentation is an effective approach to address the data scarcity issue

With LLMs (as few-shot learners), why do we still need data augmentation?

- Generic LLMs are not aware of sensitive/domain specific data (Deng et al., 2023a); e.g.,
  - Product information in a retail domain
  - Sensitive data specific to health/bank/security domain

- Utilizing LLMs for developing a product is not always possible
  - Time and compute budget constraints by small/medium scale enterprises
  - Although GPT* models are successful products, they remain proprietary and non-replicable

- LLMs themselves are used for data augmentation

LLMs are Moving the Field Forward …

Input prompt

The chatbot is … Initiate the conversation and react friendly to the user’s utterances. Talk about everyday topics for 10-15 turns and end the call.
###
AI: Hello. Did you sleep well last night?
User: Yes, I slept well. Thank you.
AI: That's good. I'm glad you slept well. Make sure not to skip any meal.
User: Okay. you eat well, too.
AI: Alright. I'll call you again later.
User: Bye.
###

Generated example

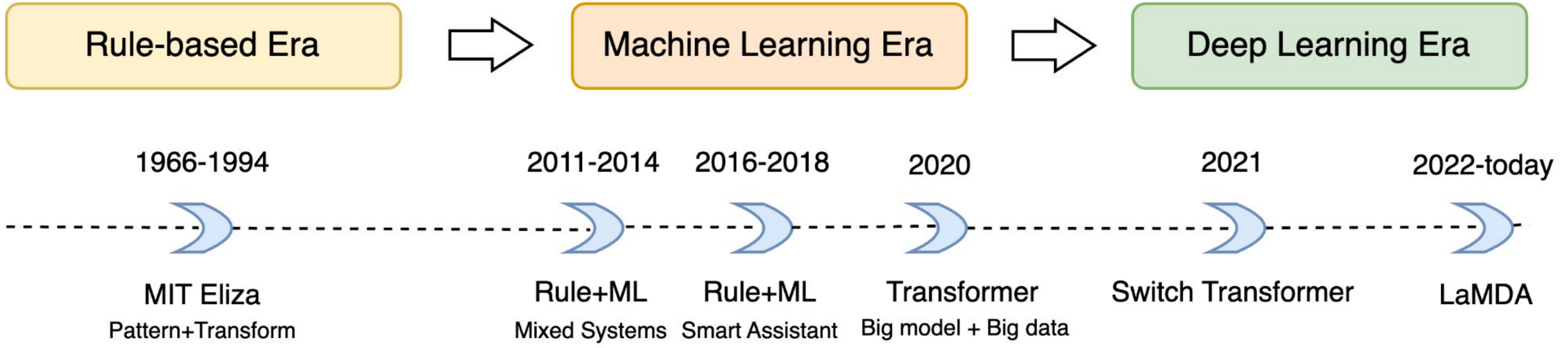
AI: Hi, how’s your health these days?
User: It’s very good.
AI: That’s good to hear. Do you work out often?
User: No, I don't have time to.
AI: I see. It would be nice if you could work out at least once a week.
User: That's true. But why do you ask?
AI: It’s just that this product has been newly released. I thought it would be helpful to you.

(Bae et al., 2022)

Introduction to **Conversational AI**

Conversational AI

Systems that are designed to simulate human-like conversations and are used for various purposes. (Zamani et al., 2023)



(Deng et al., 2023)

Types of Conversational Systems

Task Oriented

- Focuses on task completion

- Solves specific problems in a certain domain; e.g., restaurant reservation, movie ticket booking, etc.

- Developed using pipeline or end-2-end approaches

Open Domain

- Aims to chat with users without the task and domain restrictions

- Engage in conversations with users across a wide range of topics and domains

- Usually fully data-driven

Conv. Information Seeking

- Designed to assist users in seeking and retrieving information through natural language dialogue interactions.

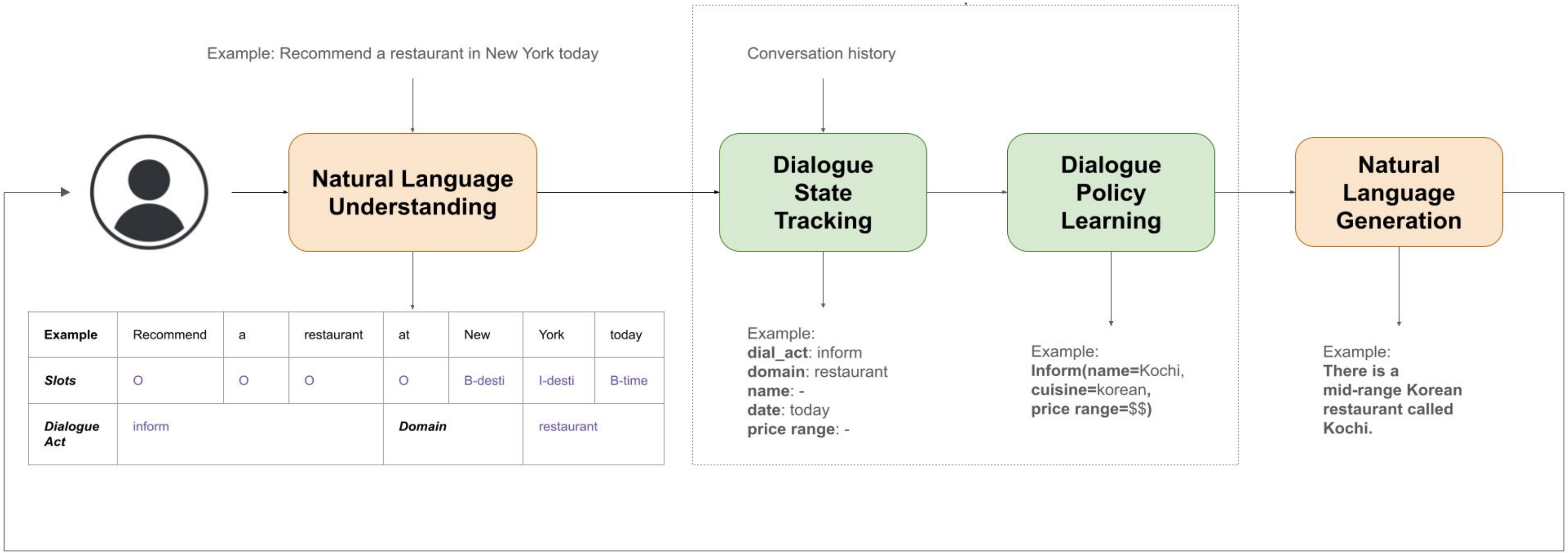
- Three main areas: conversational search, conversational (QA), and conversational recommendation

(Zamani et al., 2023), (Ni et al., 2023)

Task-oriented Dialogue Systems

- Generates responses to reach user's objectives
- Applications in various tasks such as *booking a flight, restaurant reservation, hotel recommendation, chatbots* (Fellows et al., 2021, Wen et al., 2016)
- Challenges (Kwan et al., 2023):
  - Integrating domain- and task-dependent knowledge
  - Integrating this knowledge with natural language
  - Limited data

Task-oriented Dialogue Systems



Example	Recommend	a	restaurant	at	New	York	today
Slots	O	O	O	O	B-desti	I-desti	B-time
Dialogue Act	inform			***Domain***		restaurant	

Challenges of Task Oriented Dialogue Systems

- **Cross domain transfer** (Lee et al., 2018)
    - Task-specific structural constraints make it difficult to expand to new domains
- **Diversity and coverage** (Budzianowski et al., 2018)
    - Users interact in a multitude of ways towards the same goal
- **Accuracy** (Wan et al., 2022, Yoo et al., 2020 , Terragni et al., 2023)
    - Systems need to correctly understand the state of the dialogue

Example of TOD

User: Book a restaurant in Orlando for 4 people.

System: What type of food and price range should I look for?

User: I'd like a moderately priced taiwanese restaurant.


```
"user_intents": ["BOOK_RESTAURANT"],

"system_acts": [
        { "slot": "price_range",    "type": "REQUEST" },
    { "slot": "category",         "type": "REQUEST" }],

"user_acts": [
        { "type": "INFORM" }],

"user_goal": [
    "domain": "restaurant",
    "user_intent": ["BOOK_RESTAURANT"],
    {"act": "inform",
        {"slot": "location",         "value": "orlando"},
        {"slot": "price_range",      "value": "moderately priced"},
        {"slot": "category",         "value": "taiwanese"}.
    },
    {"act": "request",
        {"slot": "price_range"},
        {"slot": "category"},
    } ]

"dialog_frame": [
    {"act": "request"},
    {"slot": "date"},
    {"slot": "time"}}]

"belief_state": [
    {"act": "inform",
        {"slot": "location",         "value": "orlando"},
        {"slot": "price_range",      "value": "moderately priced"},
        {"slot": "category",         "value": "taiwanese"}.
    },
    {"act": "request",
        {"slot": "date"},
        {"slot": "time"}
    }]
```

Open Domain Dialogue Systems

Generative Systems

Use sequence-to-sequence models to generate responses that may not be in the training corpus

Retrieval Systems

Retrieval natural and relevant pre-existing responses from a corpus

Ensemble Systems

Combine generative and retrieval methods to:

- Refine retrieved responses using generative methods **or**
- Compare retrieved and generated responses and select the best ones

Challenges of Open Domain Dialogue Systems

- **Coherence**
  - Responses are context-aware (e.g., based on conversation history)
- **Engagement**
  - Avoid making dull responses
- **Informativeness**
  - Responses are based on documents, pre-defined FAQs, and/or knowledge graphs
- **Proactivity**
  - Drive the discussion topic forward (**target-guided and policy planning**)

(Chen et al., 2023, Deng et al., 2023)

Example ODD

Crowdsourcing sample for Persona-chat dataset (Zhang et al., 2018)

Persona 1	Persona 2
I like to ski My wife does not like me anymore I have went to Mexico 4 times this year I hate Mexican food I like to eat cheetos	I am an artist I have four children I recently got a cat I enjoy walking for exercise I love watching Game of Thrones

[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

Generated dialogue form PersonaChatGen (Lee et al., 2022)

P1's Persona	P2's Persona
I have three brothers and sisters. I am the middle child. I'm a big fan of Beyoncé. I like reading books about romance. I have been diagnosed with post-traumatic stress disorder.	I love food and I love to eat. I am a woman who loves fashion. I love reality TV. I prefer to watch comedies. I have hay fever.

Generated Dialogue

P1: Hi, how are you?
P2: I'm good! I love food and I love to eat.
P1: Oh, what's your favorite food?
P2: I love all kinds of food, but my favorite food is probably sushi.
P1: Sushi is really good! I love it too.
P2: Yeah, it's one of my favorite things to eat. I also love fashion.
P1: Beyoncé is a really good fashion icon.
P2: Yeah, she is! I love her style. I also love reality TV.
P1: Oh, what's your favorite reality TV show?
P2: I love a lot of reality TV shows, but my favorite is probably The Bachelor.
P1: The Bachelor is a really good show! I love it too.
P2: Yeah, it's one of my favorite shows. I also love comedies.
P1: What's your favorite comedy?
P2: I love all kinds of comedies, but my favorite is probably The Office.
P1: The Office is a really good comedy! I love it too.
P2: Yeah, it's one of my favorite shows. I also have hay fever.

Conversational Information Seeking

- The primary objective of a CIS system is to satisfy the information needs of users
- Main areas:

Conversational search & Conversational Question Answering (QA)

find specific information, allowing users to pose multiple questions

Conversational Recommendation Systems

suggest items to users based on their previous interactions, serving as personalized information-seeking tools

(Zamani et al., 2023), (Wang et al., 2023)

Challenges of Conversational Information Seeking Systems

- **Control the generation process**
    - Topic shifting
- **Multi-evidence answer generation**
    -
- **Proactivity**
    - Mixed-initiative: Asking clarification questions

(Wu et al., 2022), (Deng et al., 2023)

Example CIS

(Feng, et al., 2020)

Dialogue turn	Dialogue act	Grounded Passage
U1: Hello, I forgot to update my address, can you help me with that?	Query	s_id: 1, text: 1. Forgetting to Update Address (A title of section)
A2: hi, you have to report any change of address to DMV within 10 days after moving. You should …	Response	s_id: 4, text: By statute , you must report a change of address to DMV within ten days of moving. That is the case for the address associated with your license …
U3: Can I do my DMV transactions online?	Query	s_id: 56, text: 5. Not Bringing Proper Documentation to DMV Office (A title of section)
A4: Yes, you can sign up for MyDMV for all the online transactions needed.	Response	s_id: 56, text: Sign up or log into MyDMV
…		
U9: Can you tell me more about Traffic points and their cost?	Query	s_id: 40, text: 4. Understanding how Much Traffic Points Cost (A title of section)
A10: Traffic points is the system used by DMV to track dangerous drivers. The cost …	Response	s_id: 41, text: DMV maintains a point system to track dangerous drivers. Often , motorists convicted …

Tutorial Agenda