

# Advanced Natural Language Processing

## CIT4230002

Prof. Dr. Georg Groh  
Yan Pan, M.Sc.

# Lecture

## Conversational AI with Domain Adaptation

- **Part 1: Conversational AI**
  - **What kinds of problem**
  - **Dialogue systems**
  - **Personal assistants today**
  - **ChatGPT**
- Part 2: Conversational question answering with domain knowledge
- Part 3: Conversational AI with domain Adaptation
- Part 4: Evaluation and BlenderBot 2.0

# What kinds of problems?

## Chitchat (social bot)

“I am smart”

“I” talk like a human

## Goal-oriented dialogues

“I have a question”

Information consumption

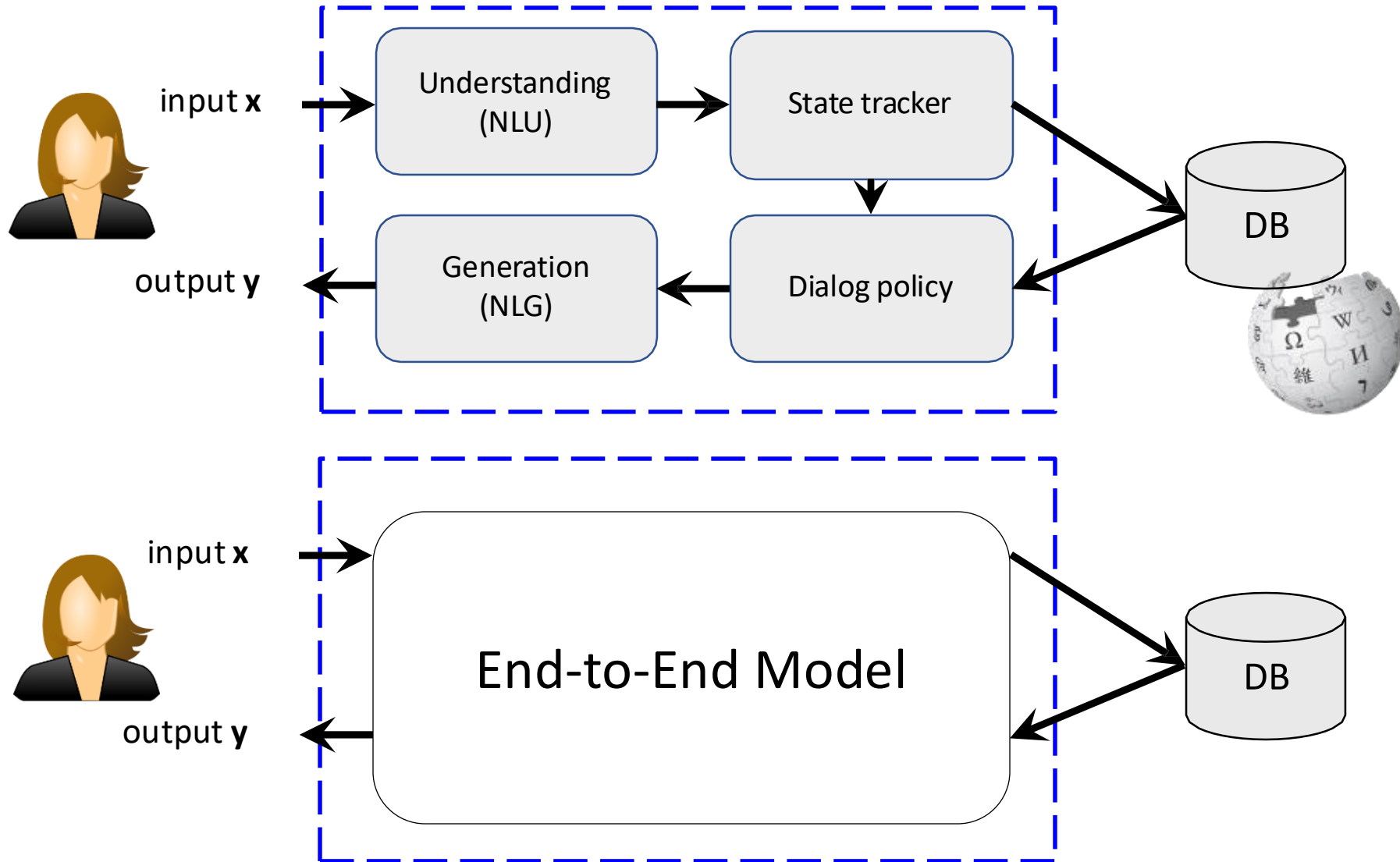
“I need to get this done”

Task completion

“What should I do?”

Decision support

# Dialog Systems



[1,2,3,4,5,6,7,8]

# Personal assistants today

## Personal assistants

Google Now

Siri

Amazon echo

## ChatGPT



Hello ChatGPT, I want to know how many people live in Munich?



Hello! As of 2021, the estimated population of Munich is around 1.57 million people. It is the third-largest city in Germany, after Berlin and Hamburg.



- ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.




What is ChatGPT?








ChatGPT is a large language model developed by OpenAI, which is designed to generate human-like responses to text-based conversations. As an AI language model, I have been trained on vast amounts of text data to be able to understand and generate natural language responses in a way that can simulate human conversation. I can answer questions, engage in discussions, and provide information on a wide range of topics. My purpose is to help people by providing quick and accurate answers to their queries or assisting them with tasks that require natural language understanding.







- The dialogue format makes it possible for ChatGPT to answer follow up questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.
- The dialogue format also supports single-turn tasks with the task descriptions.

 Hello, who are you?

 Hello! I am ChatGPT, a language model created by OpenAI. I'm designed to assist with various tasks, answer questions, and engage in conversations on a wide range of topics. How can I help you today?   

 I want to know how many people live in munich.

 As of my knowledge cutoff in September 2021, Munich, Germany had a population of approximately 1.57 million people. However, please note that population figures can change over time due to various factors such as birth rates, migration, and urban development. For the most accurate and up-to-date population information, I recommend consulting official sources or conducting a search with the current date.   



- ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers.
  - During RL training, there's currently no source of truth;
  - Training the model to be more cautious causes it to decline questions that it can answer correctly;
  - Supervised training misleads the model because the ideal answer depends on what the model knows, rather than what the human demonstrator knows.
- ChatGPT is sensitive to tweaks to the input phrasing or attempting the same prompt multiple times. For example, given one phrasing of a question, the model can claim to not know the answer, but given a slight rephrase, can answer correctly.

Task-oriented Dialogue	Prompt	<p>Continue the dialogue as a task-oriented dialogue system called SYSTEM. The answer of SYSTEM should follow the ACTION provided next while answering the USER's last utterance: ACTION: 'Train-Inform': [['choice', '12'], ['leaveat', '05:16'], ['leaveat', '16:16']], 'Train-OfferBook': [['none', 'none']]</p> <p>USER: I need train reservations from norwich to cambridge SYSTEM: I have 133 trains matching your request. Is there a specific day and time you would like to travel? USER: I'd like to leave on Monday and arrive by 18:00. SYSTEM: There are 12 trains for the day and time you request. Would you like to book it now? USER: Before booking, I would also like to know the travel time, price, and departure time please.</p>
	ChatGPT Generation	<p>SYSTEM: The travel time for the trains from Norwich to Cambridge is approximately <b>1 hour and 30 minutes. The prices for these trains range from £10 to £30</b>, and the departure times are 5:16 and 16:16. Please let me know if you would like to book a train now.</p>
	Hallucination	<p>Extrinsic - There are no information provided about neither travel time nor price range</p>
Knowledge-Grounded Open-domain Dialogue	Prompt	<p>Can we try dialogue generation? I will give you turns and you can generate the next turn, but only one. You can also consider the knowledge of "Pierre Coffin is starred in Despicable Me 2" and "Despicable Me 2 is written by Cinco Paul" for your reference in the dialogue.</p> <p>User: Do you know of any movies directed by Pierre Coffin?</p>
	ChatGPT Generation	<p>Yes, Pierre Coffin is the co-director of the <b>animated film</b> "Despicable Me 2." Cinco Paul is the film's writer.</p>
	Hallucination	<p>animated film is a factual extrinsic hallucination.</p>

Tasks	Dataset	Metric	Reference	Fine-Tuned SOTA	Zero-Shot SOTA	ChatGPT
Task-Oriented Dialogue	MultiWOZ2.2	JGA	<a href="#">Zhao et al. (2022)</a>	60.6	46.7	24.4
	MultiWOZ2.2	BLEU	<a href="#">Nekvinda and Dušek (2021)</a>	19.1	-	5.65
	MultiWOZ2.2	Inform Rate	<a href="#">Yang et al. (2021)</a>	95.7	-	71.1

- Task-specific fully-fine-tuned models outperform ChatGPT.

Model	BLEU ↑	ROUGE-L ↑	FeQA ↑ ( <a href="#">Durmus et al., 2020</a> )
ChatGPT	4.05	18.62	15.03
GPT2	11.10	30.00	26.54

Automatic evaluation results on OpenDialKG.  
The results for GPT2 are from Dziri et al. (2021)

- Human judgment: the responses from ChatGPT are of high quality with fluent response generation as well as incorporating the provided knowledge in the response.

- Part 1: Conversational AI
- **Part 2: Conversational question answering with domain knowledge**
  - Question answering (QA)
  - **Conversational question answering**
- Part 3: Conversational AI with Domain Adaptation
- Part 4: Evaluation and BlenderBot 2.0

# Question Answering (QA)

Q Will I qualify for OSAP if I'm new in Canada?

## Selected Passages from Bing

"Visit the OSAP website for application deadlines. To get OSAP, you have to be eligible. You can apply using an online form, or you can print off the application forms. If you submit a paper application, you must pay an application fee. The online application is free."

Source: <http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/how-do-i-apply-for-the-ontario-student-assistance-program-osap/>

"To be eligible to apply for financial assistance from the Ontario Student Assistance Program (OSAP), you must be a: 1 Canadian citizen; 2 Permanent resident; or 3 Protected person/convention refugee with a Protected Persons Status Document (PPSD)."

Source: <http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/who-is-eligible-for-the-ontario-student-assistance-program-osap/>

"You will not be eligible for a Canada-Ontario Integrated Student Loan, but can apply for a part-time loan through the Canada Student Loans program. There are also grants, bursaries and scholarships available for both full-time and part-time students."

Source: <http://www.campusaccess.com/financial-aid/osap.html>

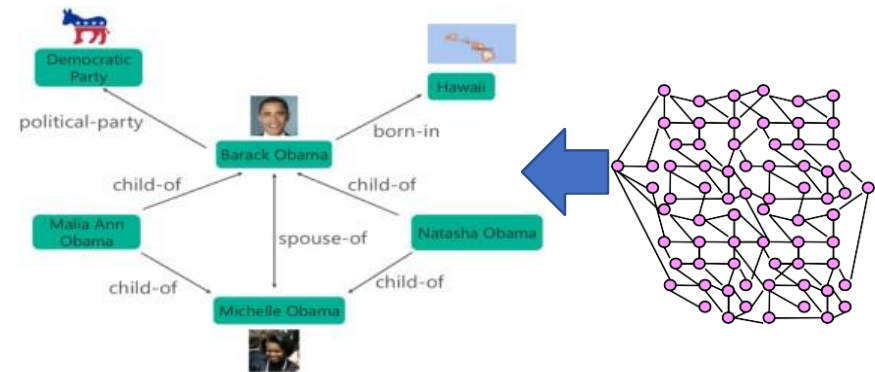
### Answer

No. You won't qualify.

Text-QA

What is Obama's citizenship?

## Selected subgraph



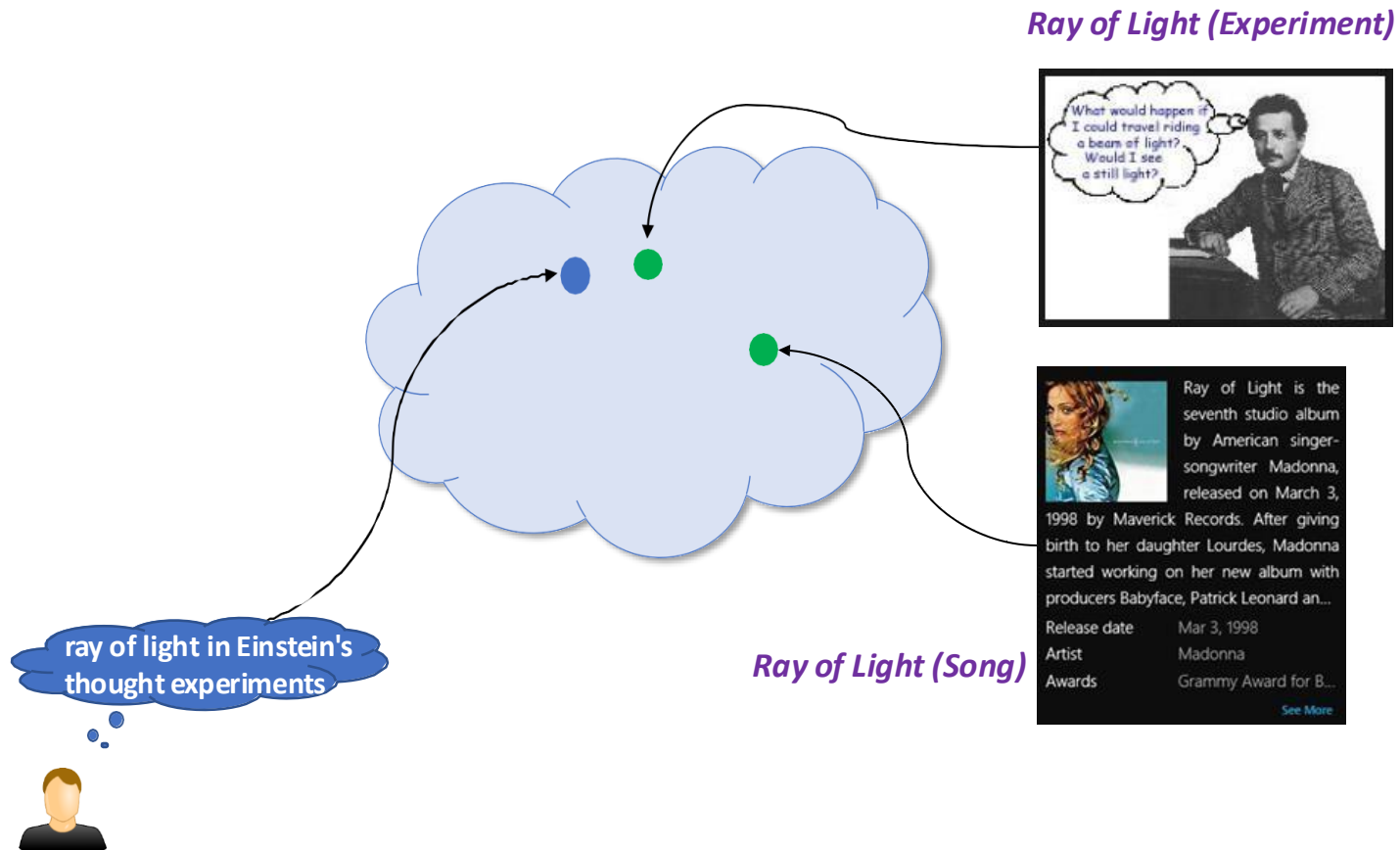
### Answer

USA

Knowledge Graph-QA

[1,2,12]

# Question Answering (QA)



- Dense Passage Retriever (DPR)
- Retrieval can be practically implemented using dense representations alone, where embeddings are learned from a small number of questions and passages by a simple dualencoder framework.
- The embedding is optimized for maximizing inner products of the question and relevant passage vectors, with an objective comparing all pairs of questions and passages in a batch.

$$\text{sim}(q, p) = E_Q(q)^\top E_P(p).$$

$$\mathcal{D} = \{\langle q_i, p_i^+, p_{i,1}^-, \dots, p_{i,n}^- \rangle\}_{i=1}^m$$

$$\begin{aligned} & L(q_i, p_i^+, p_{i,1}^-, \dots, p_{i,n}^-) \\ = & -\log \frac{e^{\text{sim}(q_i, p_i^+)}}{e^{\text{sim}(q_i, p_i^+)} + \sum_{j=1}^n e^{\text{sim}(q_i, p_{i,j}^-)}}. \end{aligned}$$

- Positive passages: the passages relevant to a question, which is given in a QA dataset
- Negative passages: all other passages in the collection, while not specified explicitly, can be viewed as irrelevant by default any random passage from the corpus.

$$\text{sim}(q, p) = E_Q(q)^\top E_P(p).$$

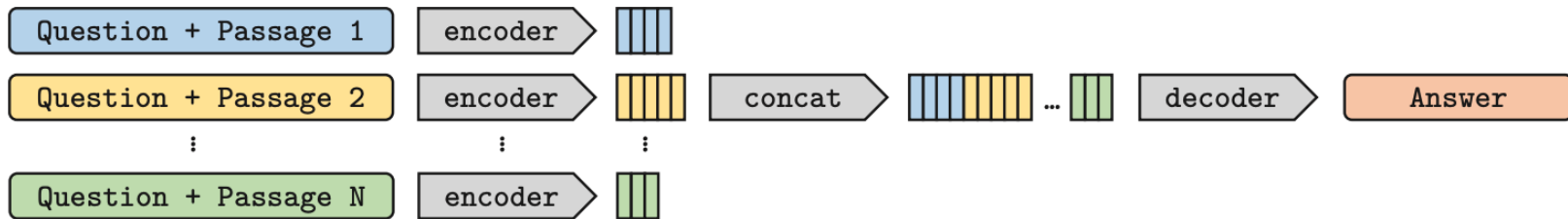
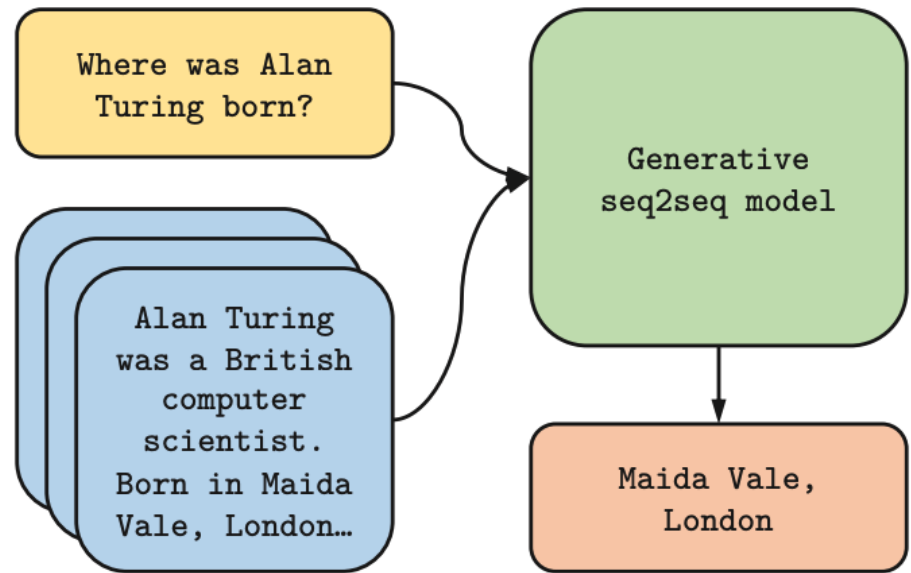
$$\mathcal{D} = \{\langle q_i, p_i^+, p_{i,1}^-, \dots, p_{i,n}^- \rangle\}_{i=1}^m$$

$$\begin{aligned} & L(q_i, p_i^+, p_{i,1}^-, \dots, p_{i,n}^-) \\ = & -\log \frac{e^{\text{sim}(q_i, p_i^+)}}{e^{\text{sim}(q_i, p_i^+)} + \sum_{j=1}^n e^{\text{sim}(q_i, p_{i,j}^-)}}. \end{aligned}$$



# Question Answering (QA)

- System retrieves support text passages from an external source of knowledge such as Wikipedia.
- A generative encoder-decoder model produces the answer, conditioned on the question and the retrieved passages.



Architecture of the Fusion-in-Decoder method

# Conversational Question Answering (CQA)

---

Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

Q<sub>1</sub>: Who had a birthday?

A<sub>1</sub>: Jessica

R<sub>1</sub>: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

Q<sub>2</sub>: How old would she be?

A<sub>2</sub>: 80

R<sub>2</sub>: she was turning 80

Q<sub>3</sub>: Did she plan to have any visitors?

A<sub>3</sub>: Yes

R<sub>3</sub>: Her granddaughter Annie was coming over

Q<sub>4</sub>: How many?

A<sub>4</sub>: Three

R<sub>4</sub>: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Q<sub>5</sub>: Who?

A<sub>5</sub>: Annie, Melanie and Josh


R<sub>5</sub>: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.


---

Conversational Question Answering (CQA): A system is required to understand the given context and then engages in multi-turn QA to satisfy a user's information needs.

CoQA

# Conversational Question Answering (CQA)

 agent

 end user

**U1** :I heard about a disability benefit for clothing. What is it about?

**A2**: This is a disability compensation benefit known as an annual clothing allowance. [2]&[3]

**U3**: I am quite interested about such benefits.

**A4**: Would you like to find out if you are eligible?

**U5**: Absolutely.

**A6**: Any damages to your clothes that were caused by prosthetic or orthopedic device or your skin medicine? [7]

**U7**: The latter just happened to me!

**A8**: Sorry to hear that. And Is it true that you need this device or skin medicine because of an injury or illness related to your military service? [8]

**U9**: True.

**A10**: You have met the requirements. [5]&[6]

**U11**: How much do I get for such benefit?

**A12**: \$817.48 [12]

**U13**: Am I still eligible for student load if I get the disability benefit?

**A14**: *Irrelevant request detected.*

Home > Disability Benefits > Eligibility > Clothing allowance

**[1]VA clothing allowance**

[2]You may be able to get money each year to help buy new clothes. [3]This is a disability compensation benefit known as an annual clothing allowance.

**[4]Can I get disability benefits from VA?**

[5]It depends on whether you meet both of the requirements listed below.

[6]Both of these must be true:

- [7]Your prosthetic or orthopedic device or your skin medicine causes damage to your clothes, and
- [8]You need this device or skin medicine because of an injury or illness related to your military service (called a service-connected condition)

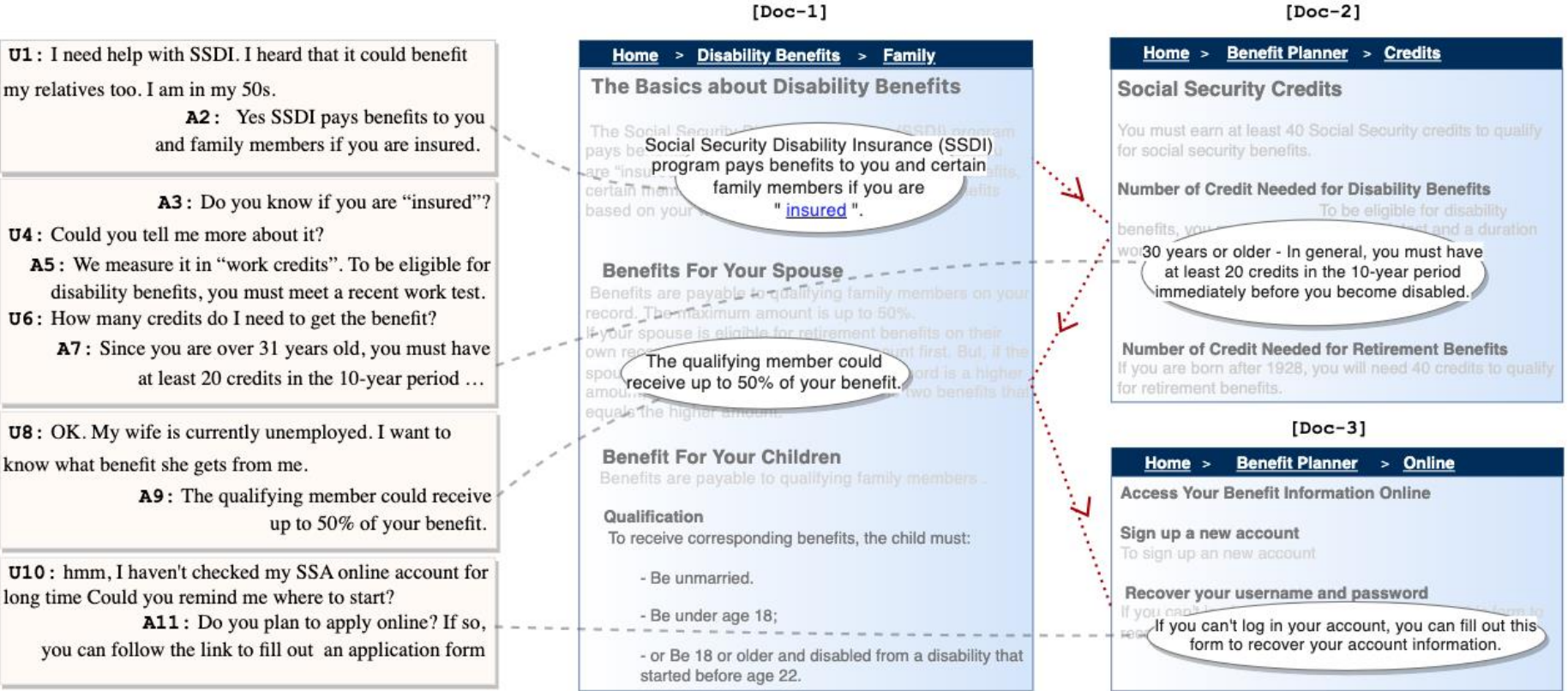
**[9]What disability benefits can I get?**

- [10]A one-time payment, or
- [11]A yearly payment

[12]Note: The current clothing allowance rate is \$817.48.

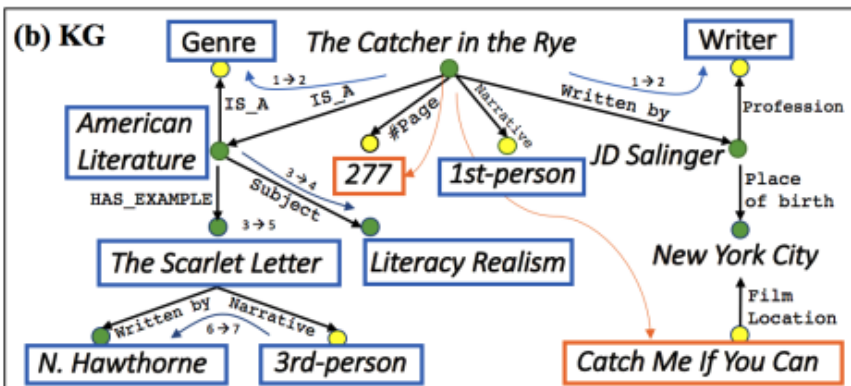
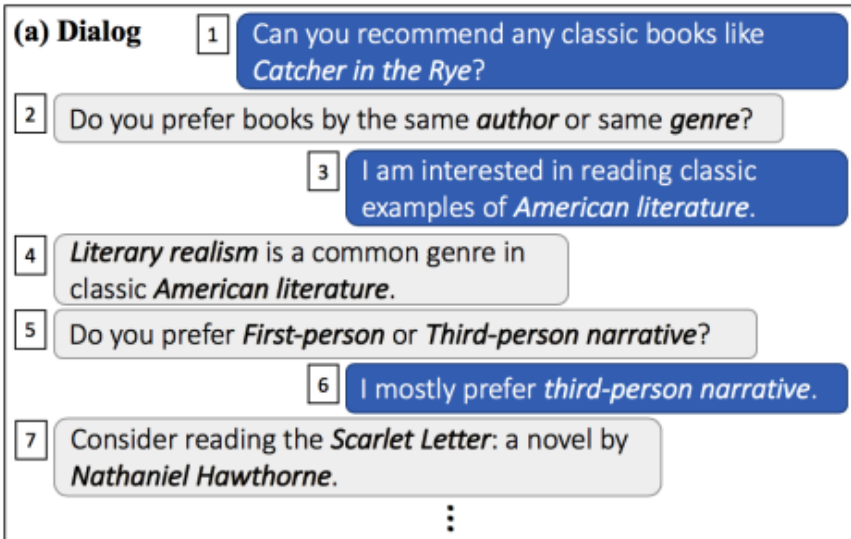
Doc2Dial dataset includes (1) a set of documents; and (2) conversations between an assisting agent and an end user, which are grounded in the associated documents.

# Conversational Question Answering (CQA)



MultiDoc2Dial:a sample goal-oriented dialogue (left) that is grounded in several documents (right).

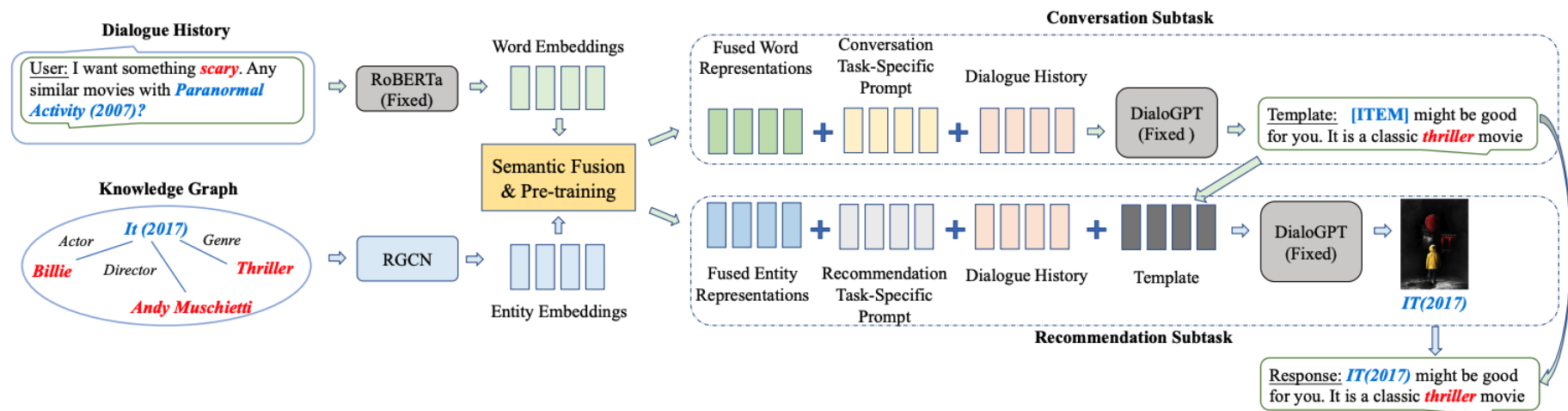
# Conversational Question Answering (CQA)



- Conversational reasoning with a parallel (a) dialog and (b) knowledge graph (KG) corpus.
- Diverse topical jumps across open-ended multi-turn dialogs are annotated and grounded with a large-scale commonfact KG.
- The walkable paths within KG can support to generate a KG entity response at each dialog turn, .

OpenDialKG

# Conversational Question Answering (CQA)

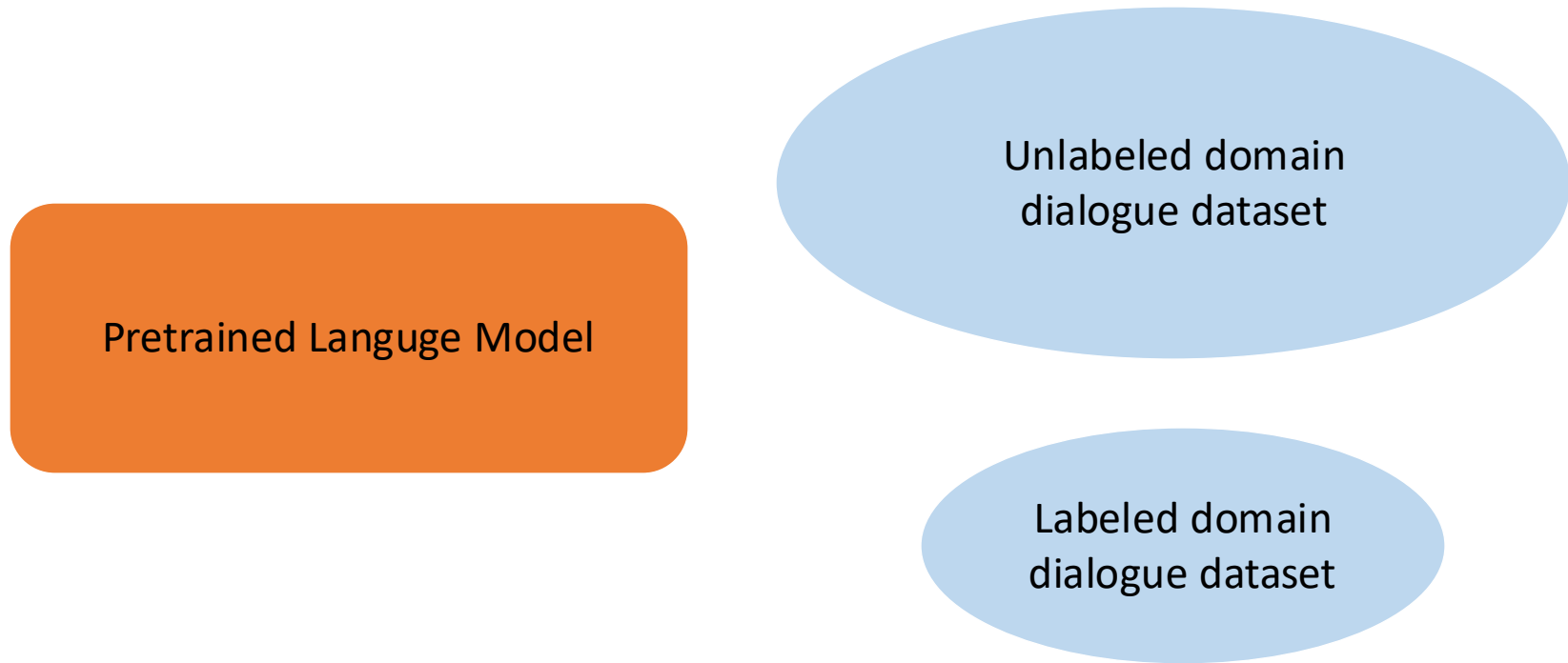


The overview of the framework UniCRS.

- Use semantic fusion and pre-training to fuse semantics from both words and entities
- Prompt the PLM to generate the response template and use the template as part of the prompt for recommendation.
- The recommended items are filled into the template as a complete response.

- Part 1: Conversational AI
- Part 2: Conversational question answering with domain knowledge
- **Part 3: Conversational AI with domain adaptation**
- Part 4: Evaluation and BlenderBot 2.0

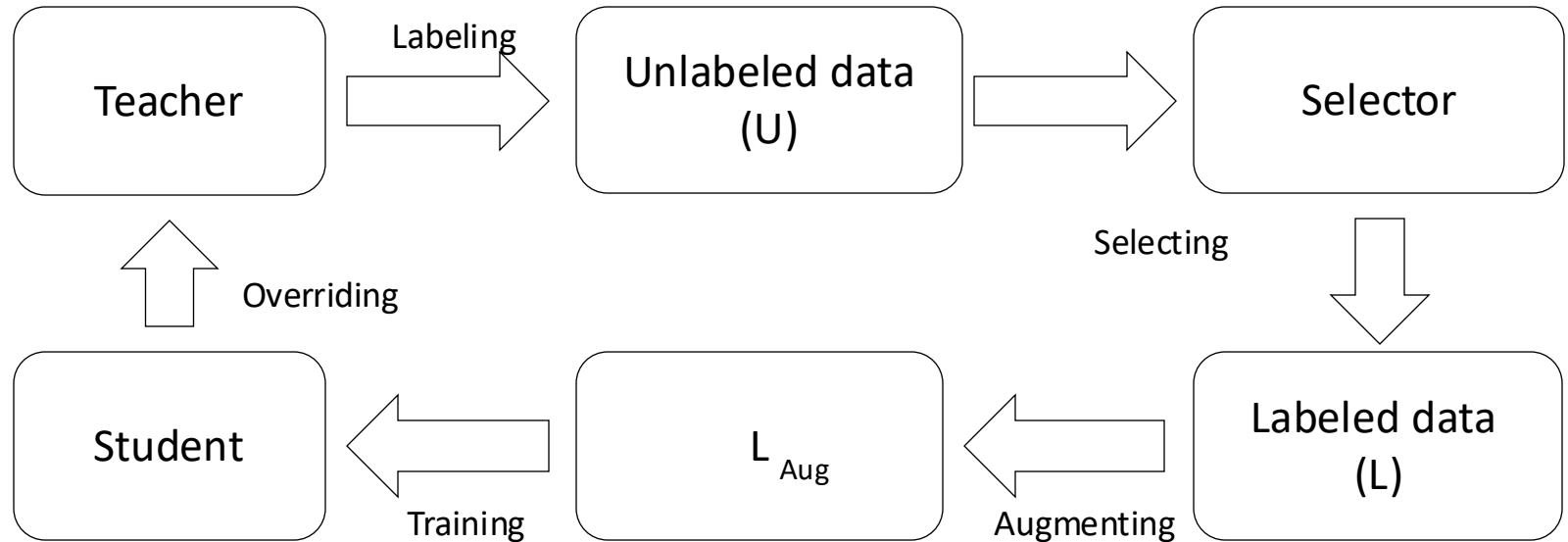
# Conversational AI with domain adaptation



- The labeling cost for different modules in task-oriented dialog systems is expensive
- Train different modules with the least amount of labeled data

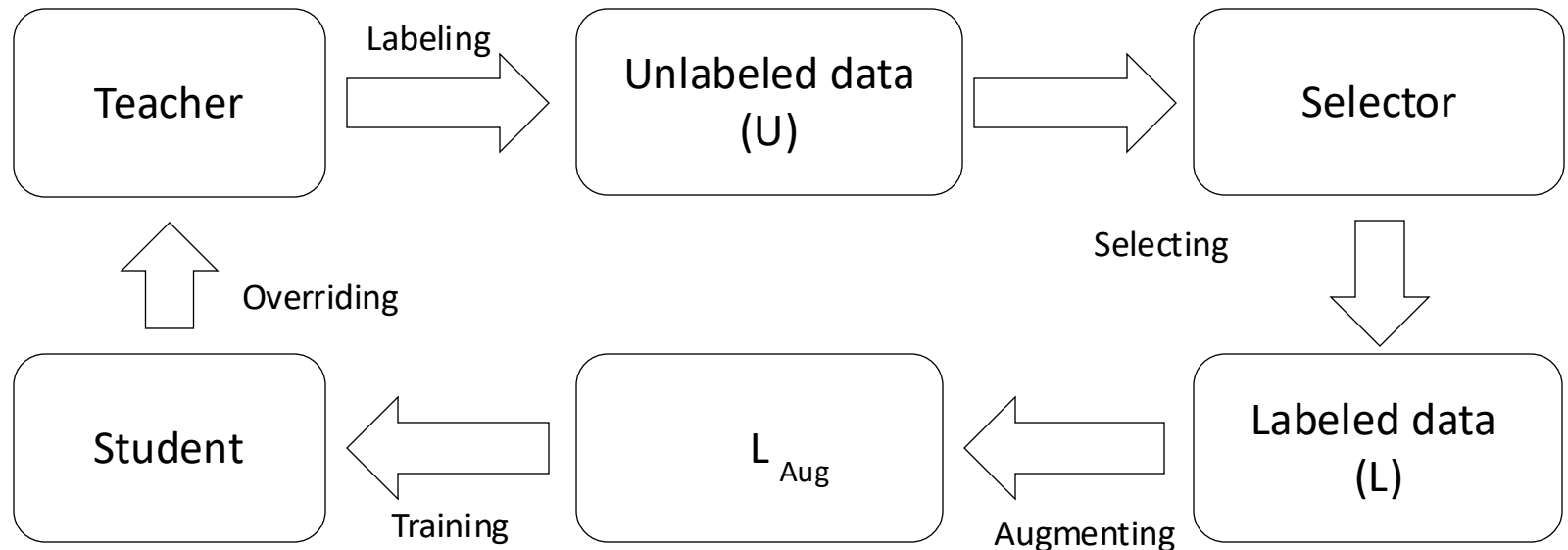


# Conversational AI with domain adaptation



- Semi-supervised self-training.
- Trains a Teacher on the labeled samples.
- The Teacher first generates predictions for data in U.
- The Selector chooses the most confident samples based on the Teacher's predictions and assign pseudo labels to them before appending to L.

# Conversational AI with domain adaptation



- L is augmented by “GradAug” to train a Student.
- GradAug first “masks” a fraction of tokens of a dialog input. Then, it reconstructs the corrupted text with a pre trained masked language model of BERT.
  - What is the status of my american airline flight ?
  - What is the status of my scheduled airline flight ?
- The trained Student becomes the Teacher in the next iteration.
- Multiple iterations are computed till the Student converges. [21]

- Part 1: Conversational AI
- Part 2: Conversational question answering
- Part 3: Conversational AI with domain adaptation
- **Part 4: Evaluation and BlenderBot 2.0**
  - Automatic evaluation
  - Human evaluation
  - BlenderBot 2.0

- **Human evaluation (crowdsourcing):**

**Context:** ... *Because of your game?*

**Input:** *Yeah, I'm on my way now*

**Response:** **Ok good luck!**

Is **this** a good<sup>1</sup> response?

	Strongly Disagree	Disagree	Unsure	Agree	Strongly Agree
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>

<sup>1</sup>: replaced as appropriate (relevant, interesting,...)

- **Automatic evaluation:**

Less expensive, but is it reliable?

- **BLEU**: n-gram overlap metric  
Machine-Translation-Based Metrics

**Reference:** John resigned yesterday .

**System:** Yesterday , John quit .

- **Distinct-N**: measures the diversity of a sentence by calculating the number of distinct unigrams and bigrams in generated responses
- **Accuracy, F1**

# Automatic evaluation

<b>Input:</b>	<i>How are you?</i>
<b>Response (gold):</b>	<b>I 'm good , thanks .</b>

Response A:	<b>Good thanks !</b>	<i><b>Semantically</b> equivalent (as in Machine Translation)</i>
Response B:	Doing pretty <b>good thanks</b>	
Response C:	Doing well thank you !	

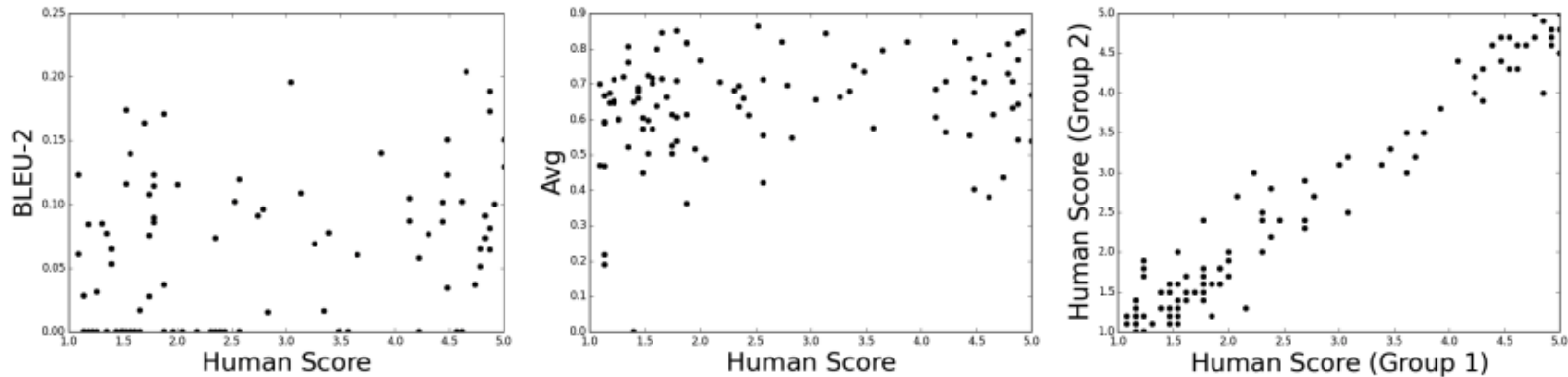
Response D:	Fantastic . How are you ?	<i><b>Pragmatically</b> appropriate</i>
Response E:	<b>I 'm</b> getting sick again .	
Response F:	Bored . you ?	
Response G:	Sleepy .	

Many false negative!

[1,2]

# Automatic evaluation

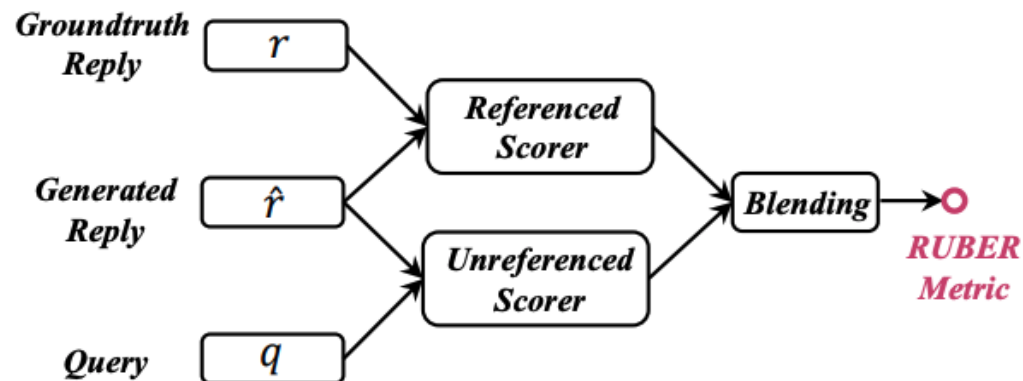
- Poor correlation with human judgments:



BLEU-2 (left), embedding average (center), and correlation between two randomly selected halves of human respondents (right)

# Automatic evaluation

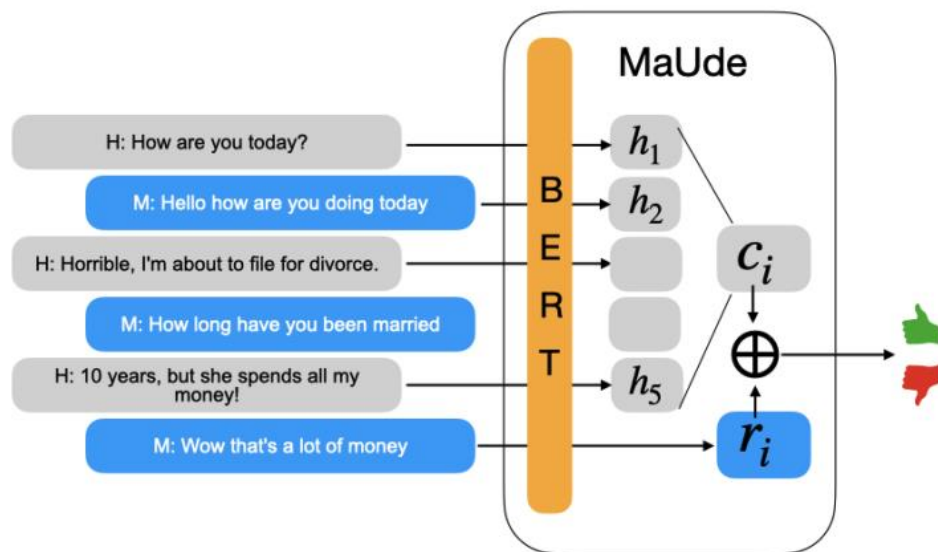
- RUBER, a Referenced metric and Unreferenced metric Blended Evaluation Routine.
  - Referenced metric measures the similarity between a generated reply and the groundtruth.
  - Unreferenced metric measures the relatedness between the generated reply and its query.





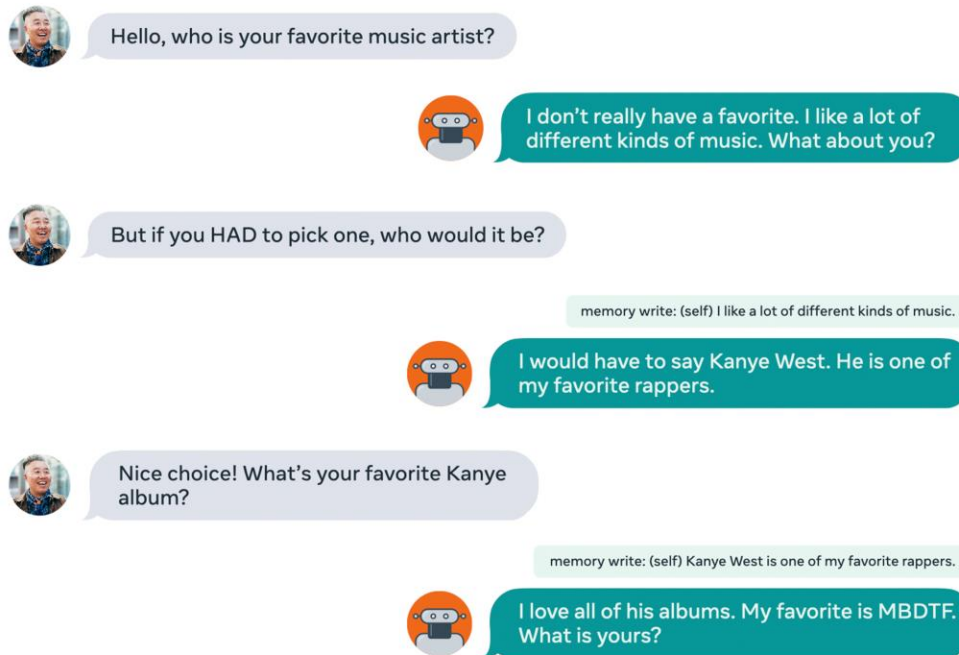
# Automatic evaluation

- MAUDE is designed to output a scalar score  $\text{score}(c_i, r_i) = R \in (0, 1)$ , which measures how appropriate a response  $r_i$  is given a dialogue context  $c_i$ .
- Differentiate between a correct response ( $\text{score}(c, r) \rightarrow 1$ ), and a negative response ( $\text{score}(c, \hat{r}) \rightarrow 0$ ), where  $\hat{r}$  represents a candidate false response for the given context  $c$ .
  - Syntactic negative samples, semantic negative samples.

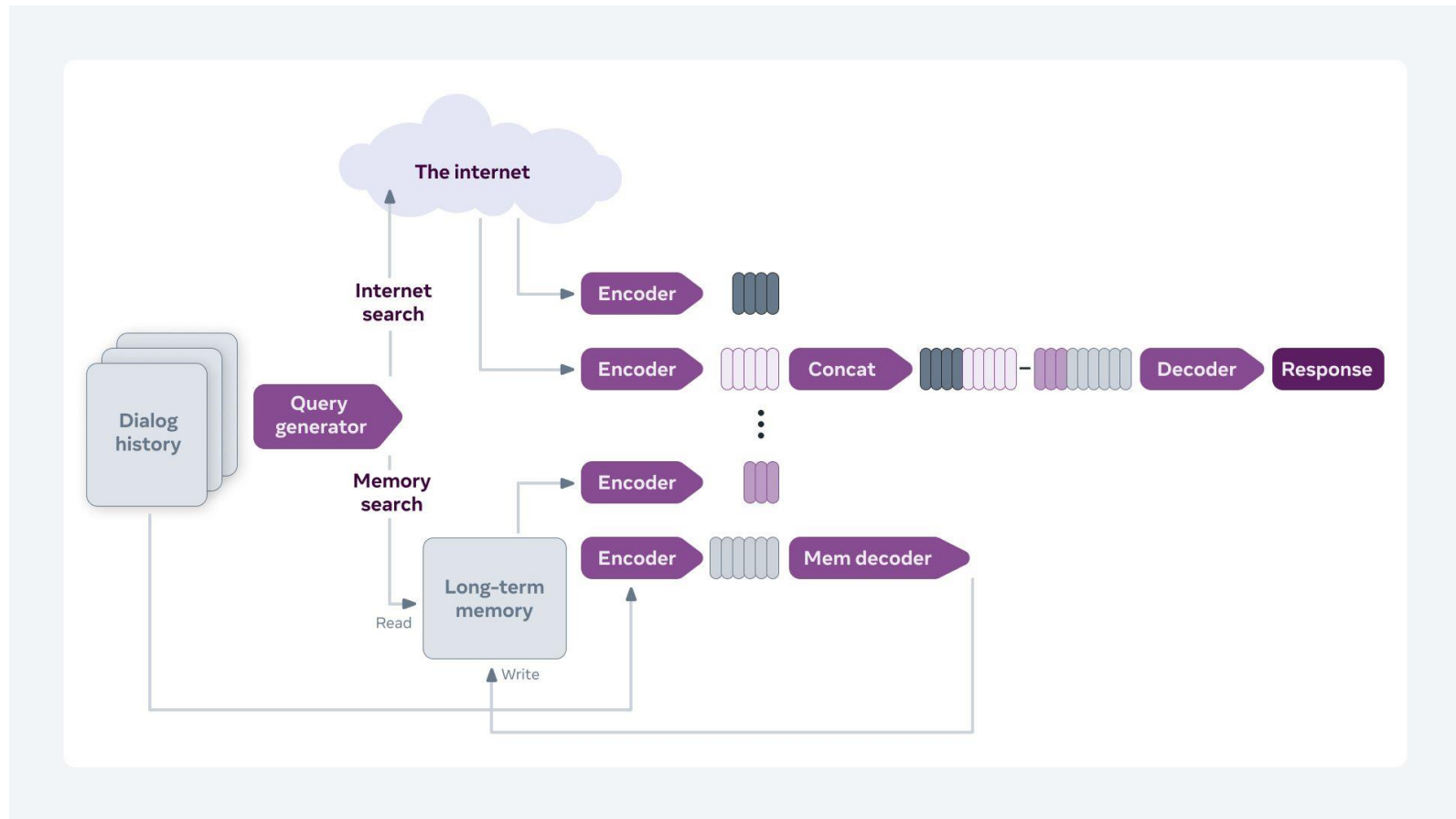


- Type of evaluators: experts, crowdsourced annotators, or end-users.
- Scale of evaluation: 1 to 5 would correspond to Very Poor, Poor, Okay, Good and Very Good. (a typical Likert scale)
- Providing a reference and a context: the output generated by the system, the context (input), and a set of reference outputs.
- Absolute v/s relative evaluation: evaluated individually or by comparing it with other outputs.
- Providing Rationale: provide reasons for their decisions.

# BlenderBot 2.0



- BlenderBot 2.0
- Long-term memory and the ability to access the internet.
- pertinent knowledge stored as a long-term memory



- The ability to access the internet:
  - generating its own search queries
  - reading the results
  - taking them into account when formulating a response.

- **Conversational AI**
- **Conversational question answering with domain knowledge**
- **Conversational AI with domain adaptation**
- **Evaluation and BlenderBot 2.0**

## Minimal

- Work with the slides

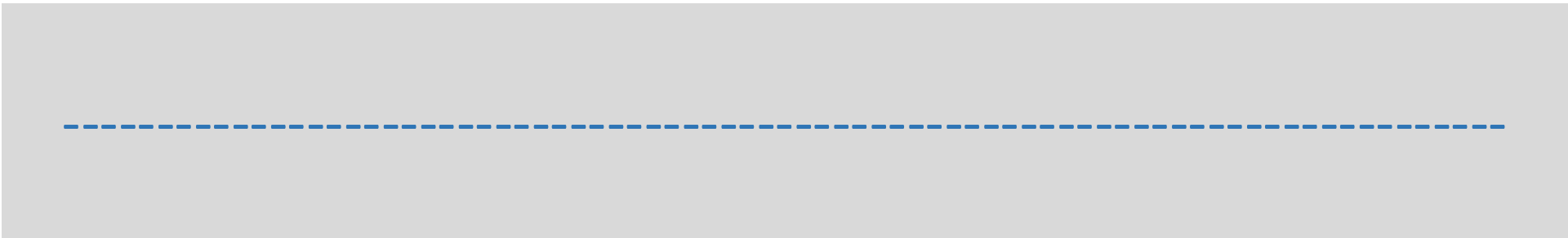
## Standard

- Minimal approach + read reference 13

## In-Depth

- = standard approach + read reference 25

See you next time!





- (1) Jianfeng Gao and Michel Galley. 2019. Neural approaches to conversational ai - tutorial at icml 2018. 191 In ICML 2019 Tutorial
- (2) Gao, J., Galley, M., & Li, L. (2018, June). Neural approaches to conversational AI. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 1371-1374).
- (3) S. Young, M. Gašić, B. Thomson and J. D. Williams, "POMDP-Based Statistical Spoken Dialog Systems: A Review," in *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1160-1179, May 2013, doi: 10.1109/JPROC.2012.2225812.
- (4) Tur, G., & De Mori, R. (2011). *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.
- (5) Alan Ritter, Colin Cherry, and William B. Dolan. 2011. [Data-Driven Response Generation in Social Media](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, Edinburgh, Scotland, UK.. Association for Computational Linguistics.
- (6) Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. [A Neural Network Approach to Context-Sensitive Generation of Conversational Responses](#). In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

# Bibliography

- (7) Vinyals, O., & Le, Q. V. (2015). A Neural Conversational Model. *arXiv preprint arXiv:1506.05869*.
- (8) Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. [Neural Responding Machine for Short-Text Conversation](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.
- (9) OpenAI (2022). Introducing ChatGPT. URL: <https://openai.com/blog/chatgpt>
- (10) Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., ... & Fung, P. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.
- (11) Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. 2021. Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- (12) Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., ... & Wang, T. (2016). Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.
- (13) Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W. T. (2020, November). Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769-6781).

- (14) Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 874–880, Online. Association for Computational Linguistics.
- (15) Reddy, S., Chen, D., & Manning, C. D. (2019). Coqa: A conversational question answering challenge. Transactions of the Association for Computational Linguistics, 7, 249-266.
- (16) Zaib, M., Zhang, W. E., Sheng, Q. Z., Mahmood, A., & Zhang, Y. (2022). Conversational question answering: A survey. Knowledge and Information Systems, 64(12), 3151-3195.
- (17) Feng, S., Wan, H., Gunasekara, C., Patel, S. S., Joshi, S., & Lastras, L. (2020, November). doc2dial: A Goal-Oriented Document-Grounded Dialogue Dataset. In Conference on Empirical Methods in Natural Language Processing.
- (18) Feng, S., Patel, S. S., Wan, H., & Joshi, S. (2021, November). MultiDoc2Dial: Modeling Dialogues Grounded in Multiple Documents. In Conference on Empirical Methods in Natural Language Processing.
- (19) Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 845–854, Florence, Italy. Association for Computational Linguistics.
- (20) Wang, X., Zhou, K., Wen, J. R., & Zhao, W. X. (2022, August). Towards Unified Conversational Recommender Systems via Knowledge-Enhanced Prompt Learning. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 1929-1937).

- (21) Fei Mi, Wanhao Zhou, Lingjing Kong, Fengyu Cai, Minlie Huang, and Boi Faltings. 2021. Self-training Improves Pre-training for Few-shot Learning in Task-oriented Dialog Systems. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1887–1898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- (22) Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- (23) Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 110–119, San Diego, California. Association for Computational Linguistics.
- (24) Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.
- (25) Tao, C., Mou, L., Zhao, D., & Yan, R. (2018, April). Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

- (26) Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. 2020. Learning an Unreferenced Metric for Online Dialogue Evaluation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2430–2441, Online. Association for Computational Linguistics.
- (27) Sai, A. B., Mohankumar, A. K., & Khapra, M. M. (2022). A survey of evaluation metrics used for NLG systems. *ACM Computing Surveys (CSUR)*, 55(2), 1-39.
- (28) Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.
- (29) Weston, J., Shuster, K. (2021) Blenderbot 2.0, <https://ai.meta.com/blog/blender-bot-2-an-open-source-chatbot-that-builds-long-term-memory-and-searches-the-internet/>
- (30) Komeili, M., Shuster, K., & Weston, J. (2022, May). Internet-Augmented Dialogue Generation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 8460-8478).
- (31) Xu, J., Szlam, A., & Weston, J. (2022, May). Beyond Goldfish Memory: Long-Term Open-Domain Conversation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 5180-5197).