Evaluating Coherence in Dialogue Systems using Entailment

Nouha Dziri

Rasa Developer Summit

Sept 24th, 2019

UNIVERSITY OF ALBERTA

💬 Dialogue System

Goal-oriented 🎯

- Designed for short conversations.
- Accomplishes a specific task.
- Search space is narrow.

Open-ended 🗣️

- Designed for extended conversations.
- Chit-chat with humans in an open-domain context.
- Search space is big.

💬 Dialogue System

Goal-oriented 🎯

- Designed for short conversations.
- Accomplishes a specific task.
- Search space is narrow.

Evaluation is done via human-generated judgment like a **task completion test** or **user satisfaction score**.

Open-ended 🗣

- Designed for extended conversations.
- Chit-chat with humans in an open-domain context.
- Huge search space.

💬 Dialogue System

Goal-oriented 🎯

Open-ended 🗣

- Designed for short conversations.
- Accomplishes a specific task.
- Search space is narrow.

- Designed for extended conversations.
- Chit-chat with humans in an open-domain context.
- Huge search space.

It is **unclear** how to define a metric that can account comprehensively for **overall quality.**

Evaluation is hard 🤯

- Word-overlap metrics (e.g., BLEU, METEOR, ROUGE)
- Statistical (e.g., perplexity)
- Human Evaluation
- Learned Evaluation
  - ADEM [Lowe et al., ACL’17]
  - Re-evaluating ADEM [Sai et al., AAAI’19]

We would like to have a well-designed automated metric that provides an accurate evaluation of the system without any human intervention! 🦄

Dialogue quality aspects 🤔

Conversational logic can be modeled as a set of maxims, known as Grice’s maxims [Grice, “*Logic and conversation*”, 1975]:

1. **Maxim of quantity**
2. **Maxim of quality**
3. **Maxim of relevance**
4. **Maxim of manner**

Dialogue quality aspects

1. **Maxim of quantity** where one tries to be as *informative* as one possibly can, and gives as much information as is needed, and no more.

Definition taken from: https://www.sas.upenn.edu/~haroldfs/dravling/grice.html

Dialogue quality aspects

1. **Maxim of quantity** where one tries to be as *informative* as one possibly can, and gives as much information as is needed, and no more.
2. **Maxim of quality** where one tries to be *truthful*, and does not give information that is false or that is not supported by evidence.

Definition taken from: https://www.sas.upenn.edu/~haroldfs/dravling/grice.html

Dialogue quality aspects

1. **Maxim of quantity** where one tries to be as *informative* as one possibly can, and gives as much information as is needed, and no more.
2. **Maxim of quality** where one tries to be *truthful*, and does not give information that is false or that is not supported by evidence.
3. **Maxim of relevance** where one tries to be *relevant*, and says things that are pertinent to the discussion.

Definition taken from: https://www.sas.upenn.edu/~haroldfs/dravling/grice.html

Dialogue quality aspects

1. **Maxim of quantity** where one tries to be as *informative* as one possibly can, and gives as much information as is needed, and no more.
2. **Maxim of quality** where one tries to be *truthful*, and does not give information that is false or that is not supported by evidence.
3. **Maxim of relevance** where one tries to be *relevant*, and says things that are pertinent to the discussion.
4. **Maxim of manner** where one tries to be as *clear*, as *brief*, and as orderly as one can in what one says, and where one avoids obscurity and ambiguity.

Definition taken from: https://www.sas.upenn.edu/~haroldfs/dravling/grice.html

Dialogue quality (a different angle) 🤔

Control generation based on the following aspects [See et al., *"What makes a good conversation? How controllable attributes affect human judgments"*, NAACL'19]

1. **Repetition**
2. **Specificity**
3. **Response-relatedness**
4. **Question-asking**

Dialogue quality aspects (a different angle)

Control generation based on the following aspects [See et al., *"What makes a good conversation? How controllable attributes affect human judgments"*, NAACL'19]

1. **Repetition**
2. **Specificity**
3. **Response-relatedness**
4. **Question-asking**

Maxim of manner

Dialogue quality aspects (a different angle)

Control generation based on the following aspects [See et al., *“What makes a good conversation? How controllable attributes affect human judgments”*, NAACL’19]

1. **Repetition**
2. **Specificity**
3. **Response-relatedness**
4. **Question-asking**

Maxim of manner

Maxim of relevance

Dialogue quality aspects (a different angle)

Control generation based on the following aspects [See et al., *“What makes a good conversation? How controllable attributes affect human judgments”*, NAACL’19]

1. **Repetition**
2. **Specificity**
3. **Response-relatedness**
4. **Question-asking**

Maxim of manner

Maxim of relevance

Engagingness

Dialogue Consistency 🧐

The responses must be

- Self-consistent: NOT contradicting one’s previous utterances
- Aligned with the conversation history
- Tied to external knowledge or commonsense

Maxim of quality

Dialogue Consistency

🤖 I like Captain America and Star Wars.

👩 What superpowers did you awake with?

🤖 I do not like superpowers

✗ **Contradiction** 😦

Dialogue Consistency as NLI

[Dziri et al., *“Evaluating Coherence in Dialogue Systems using Entailment”*, NAACL’19]

🤖 I like Captain America and Star Wars.

👩 What superpowers did you awake with?

Premise

🤖

Moving objects with my mind — **Hypothesis 1**

I don’t know — **Hypothesis 2**

I do not like superpowers — **Hypothesis 3**

Dialogue Consistency as NLI

[Dziri et al., *“Evaluating Coherence in Dialogue Systems using Entailment”*, NAACL’19]

🤖 I like Captain America and Star Wars.

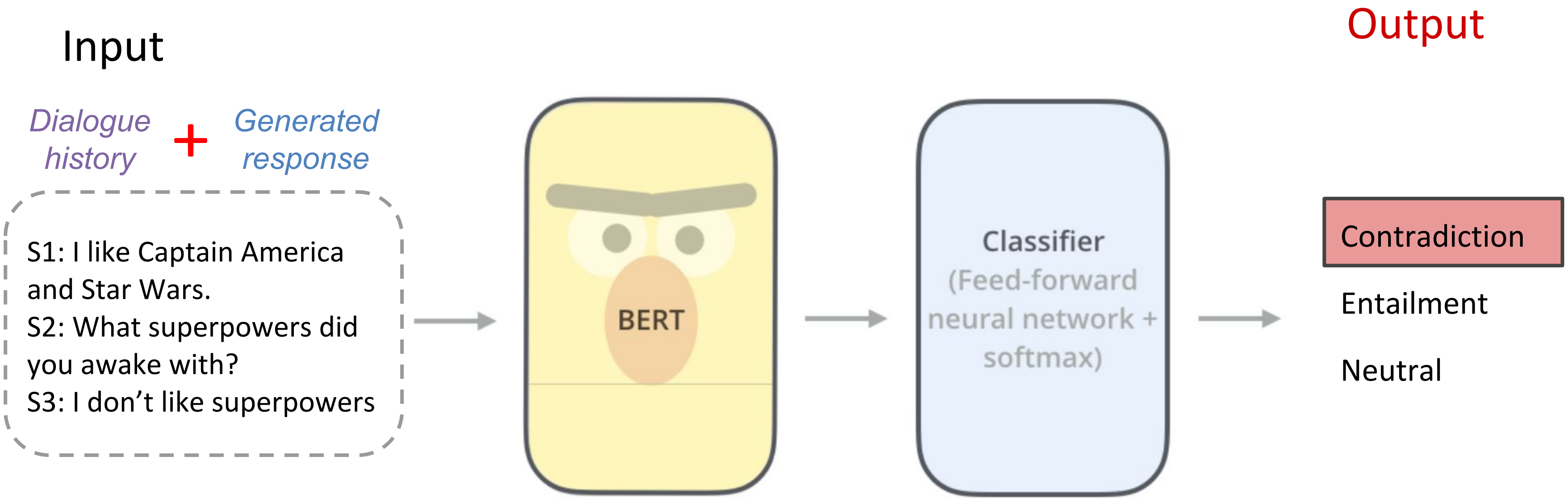
👩 What superpowers did you awake with?

🤖

Moving objects with my mind	**(Entailment)**
I don’t know	**(Neutral)**
I do not like superpowers	**(Contradiction)**

Dialogue Consistency as NLI

[Dziri et al., *"Evaluating Coherence in Dialogue Systems using Entailment"*, NAACL'19]



Input

Dialogue history + *Generated response*

S1: I like Captain America and Star Wars.
S2: What superpowers did you awake with?
S3: I don't like superpowers

Output

Contradiction

Entailment

Neutral

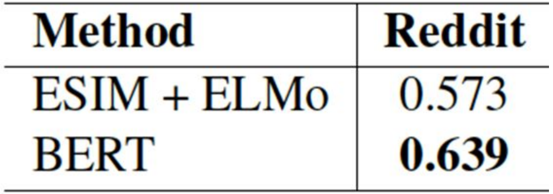
Image credit https://jalammar.github.io/illustrated-bert/

Consistency Corpus 📚

- Build a synthesized Inference Corpus based on the Persona-Chat conversational data [Zhang et al., *"Personalizing Dialogue Agents: I have a dog, do you have pets too?"*, ACL'18].
  - Natural response *as entailment*
  - Random utterances or generic responses *as neutral*
  - Grammatically-impaired utterances or contradictory examples from MNLI *as contradiction*
- Dialogue Natural Language Inference [Welleck et al., *"Dialogue Natural Language Inference"*, ACL'19]

Experiments 🧪

- Trained neural dialogue systems on a conversational dataset derived from Reddit [Dziri et al., *“Augmenting Neural Response Generation with Context-Aware Topical Attention”*, NLP4ConvAI’19].
- The model achieved an accuracy of 0.63.



Method	**Reddit**
ESIM + ELMo	0.573
BERT	**0.639**

Take-away messages

- Evaluating dialogue systems is far from being solved, researchers are still on the quest for a strong and reliable metric that highly conforms with human judgment.
- Consistency is key in evaluating dialog systems.
- Entailment techniques lay the foundations of future works to evaluate better the consistency in dialogues.

Thank you !

Questions?