Conversational Assistance Track

Summary

The focus of this track is to perform advanced research on conversational search systems and to create a reusable benchmark for open-domain information-centric conversational dialogues(http://www.treccast.ai/). There are many conversational assistant systems and despite their ability to perform simple well-defined actions, the potential to support conversational information seeking is still very limited. Several key properties defined for CAsT topics are:

- Complexity - Requires multiple turns to address the different aspects
- Diversity - Covers all domains of information topics (news, travel, health, politics, history, science, etc.)
- Answerable - Most turns should be answerable with relevant content
- Multi-source - Requires content from multiple information sources and not a single article
- Varied discourse - Varying types of conversational structural patterns

Task

The first year of this track was in 2019. The aim is to focus on candidate information ranking in context:

- Read the dialogue context: To track the evolution of the information need in the conversation, by identification of salient information needed for the current turn in the conversation.
- Retrieve Candidate Response Information: Retrieval performance over a large collection of paragraphs and identify relevant information.

Dataset

Dataset is a combination of three different text collections that mirrors major verticals for conversational agents. The goal is to retrieve passages from target open-domain text collections. The passages must be retrieved from one of the three following collections:

- English Wikipedia -
    - TREC Complex Answer Retrieval v2.1 (may increase to 3.0 pending availability).
    - Article paragraph content from Wikipedia (Wikipedia dump from December 20, 2016).
    - Approximately 5 Million articles. Publicly available.
- MS MARCO web passage data -
    - MS MARCO Passage Reranking data

    - 10 million answer candidates from Bing search. Publicly available.
- Washington Post news collection -
    - TREC Washington Post Corpus used by the TREC News Track
    - 608,180 news articles and blog posts from 2012 through 2017
    - The collection requires a data license agreement.

Training topics

Sample training topics are provided from two sources. The first is manually constructed dialogues and the second is derived from MARCO web search session data.

Assessing guidelines

Response pooling of the top responses for systems is performed across the participants. Responses are assessed on a five-point graded relevance scale: Fails to meet, Slightly meets, Moderately meets, Highly meets, and Fully meets.

- 4 Fully Meets - Passage is the 'perfect' single response to the utterance
- 3 Highly meets - Passage answers the utterance and is focused on the answer i.e., what a voice assistant should deliver.
- 2 Moderately Meets - The passage answers the utterance, but is focused on something related i.e., it might initially be clear why a voice assistant picked this passage.
- 1 Slightly meets - The answer can be inferred from the passage with reasonable effort i.e., better than nothing
- 0 Fails to meet - Not relevant

Evaluation of Runs

- Ranking depth is the same as for adhoc search, but the focus is on the earlier positions (1, 3, 5) for the conversational scenario.
- Standard ranking metrics such as P@1, P@3, ERR and MAP are calculated using the judgments.

Approach to the task

Sequential models like LSTM and Attention-based models can be used for the conversation given contextual information.