TopiOCQA: Open-domain Conversational Question Answering with Topic Switching

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, Siva Reddy

Motivation

Current conversational question-answering datasets are limiting in two ways:

- They do not contain **topic switches**.
- They are not **open-domain**.

We introduce TopiOCQA, an open-domain conversational question-answering dataset with topic switches based on Wikipedia corpus of 5.9 million documents.

Example

Q_1: when was the byzantine empire born what was it originally called?
A_1: 5th century AD and was called Eastern Roman Empire, or Byzantium
Topic: Byzantine Empire

. . . . .

Q_3: which battle or event marked the fall of this empire?
A_3: A six-year-long civil war followed by attack from Sultan Mehmed's army
Topic: Byzantine Empire

Q_4: did he conquer other territories as well?
A_4: Yes. Anatolia and in Southeast Europe as far west as Bosnia
Topic: Mehmed the Conqueror

Q_5: where is the first area located in present day terms?
A_5: Turkey
Topic: Anatolia

. . . . .

Q_9: were any of these cities associated with the first empire you were discussing?
A_9: The Ottomans made the city of Ankara . . . Anatolia Eyalet and then the Angora Vilayet
Topic: Ankara

Experiments

We evaluate TopiOCQA on QA models with following question representations

Q_1: Who is lead singer of Rage Against the Machine?
A_1: Zack de la Rocha
Q_2: When was it formed?
A_2: 1991
Q_3: Was it nominated for any award?

AllHistory: Who is lead singer of Rage Against the Machine [SEP] Zack de la Rocha [SEP] When was it formed [SEP] 1991 [SEP] Was it nominated for any award?
Rewrites: Was Rage Against the Machine nominated for any award?

Model types evaluated:

- **open-book** – models have access to document corpus. Dense Passage Retrieval (DPR, extractive) and Fusion-in-Decoder (FiD, abstractive)
- **closed-book** – models don't have access to document corpus. GPT-3.

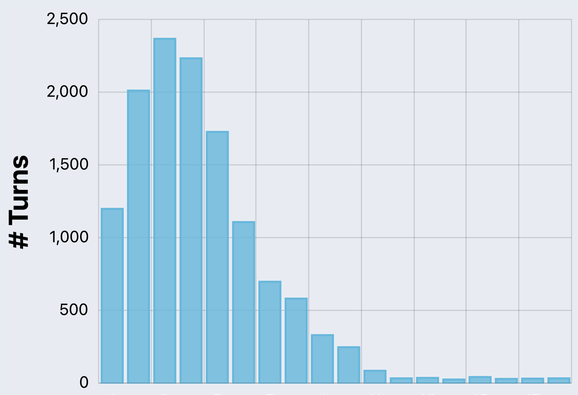
In **Ideal Retriever**, reader gets the gold passage, instead of retrieved passages.

Model	Question Rep	Trained Retriever		Ideal Retriever	
		EM	F1	EM	F1
Human		**40.2**	**70.1**	–	–
DPR	AllHistory	21.0	43.4	29.7	54.2
	Rewrites	17.2	36.4	29.8	53.8
FiD	AllHistory	33.0	55.3	38.3	65.5
	Rewrites	23.5	44.2	34.5	61.9
GPT-3		12.4	33.4	–	–

Passage retrieval is a significant bottleneck for the task.

Topic Switching Analysis

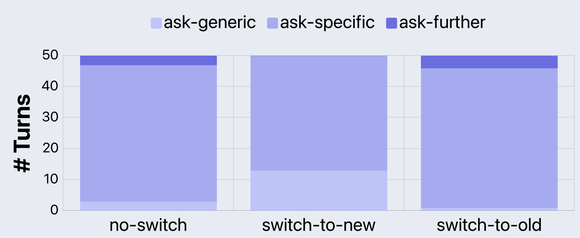
Q1. How long does the topic remain the same?



A1. 3-4 turns

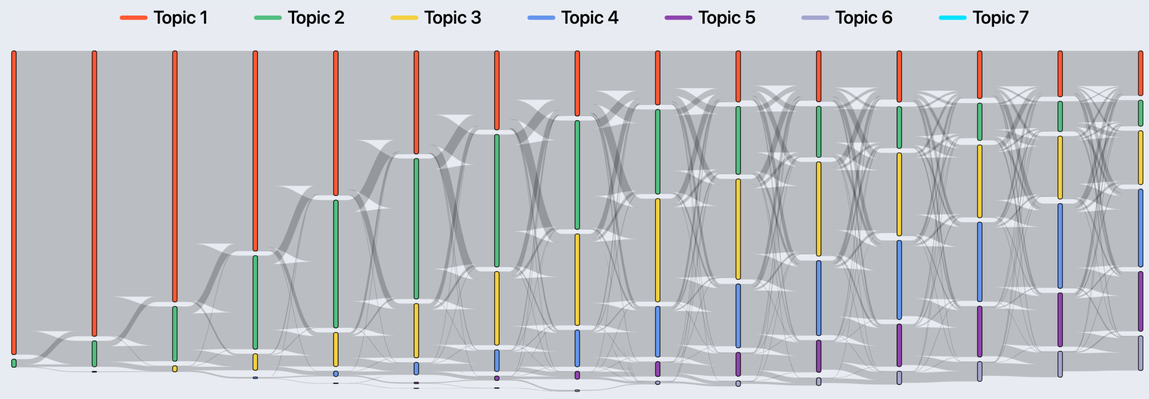
Q2. Does topic switching influence the type of question?

- `ask-generic` - *Who is Maria*
- `ask-specific` - *Is she married?*
- `ask-further` - *To whom?*



A2. Yes, generic questions are expected more at a topic switch.

Q3. How do topics switch within a conversation?



A3. Earlier turns introduce new topics, latter turns have complex interactions.

Statistics

# Turns	50,466
# Conversations	3920
# Tokens / Question	6.92
# Tokens / Answer	11.75
# Turns / conversation	13
# Topics / conversation	4

For code, models, and dataset explorer, visit
mcgill-nlp.github.io/topiocqa