University of Massachusetts Amherst BE REVOLUTIONARY

TopicGPT: A Prompt-based Framework for Topic Modeling

Chau Minh Pham[1], Alexander Hoyle[2], Simeng Sun[1], Mohit Iyyer[1]

[1]University of Massachusetts Amherst, [2]University of Maryland

Code: https://github.com/chtmp223/topicGPT

Introduction

- Conventional topic models represent topics in a **bag-of-words** format that often requires "reading the tea leaves" to interpret; additionally, they offer minimal **semantic control** over topics.
- We introduce **TopicGPT**, a framework that uses **large language models (LLMs)** to uncover latent topics in text corpora.

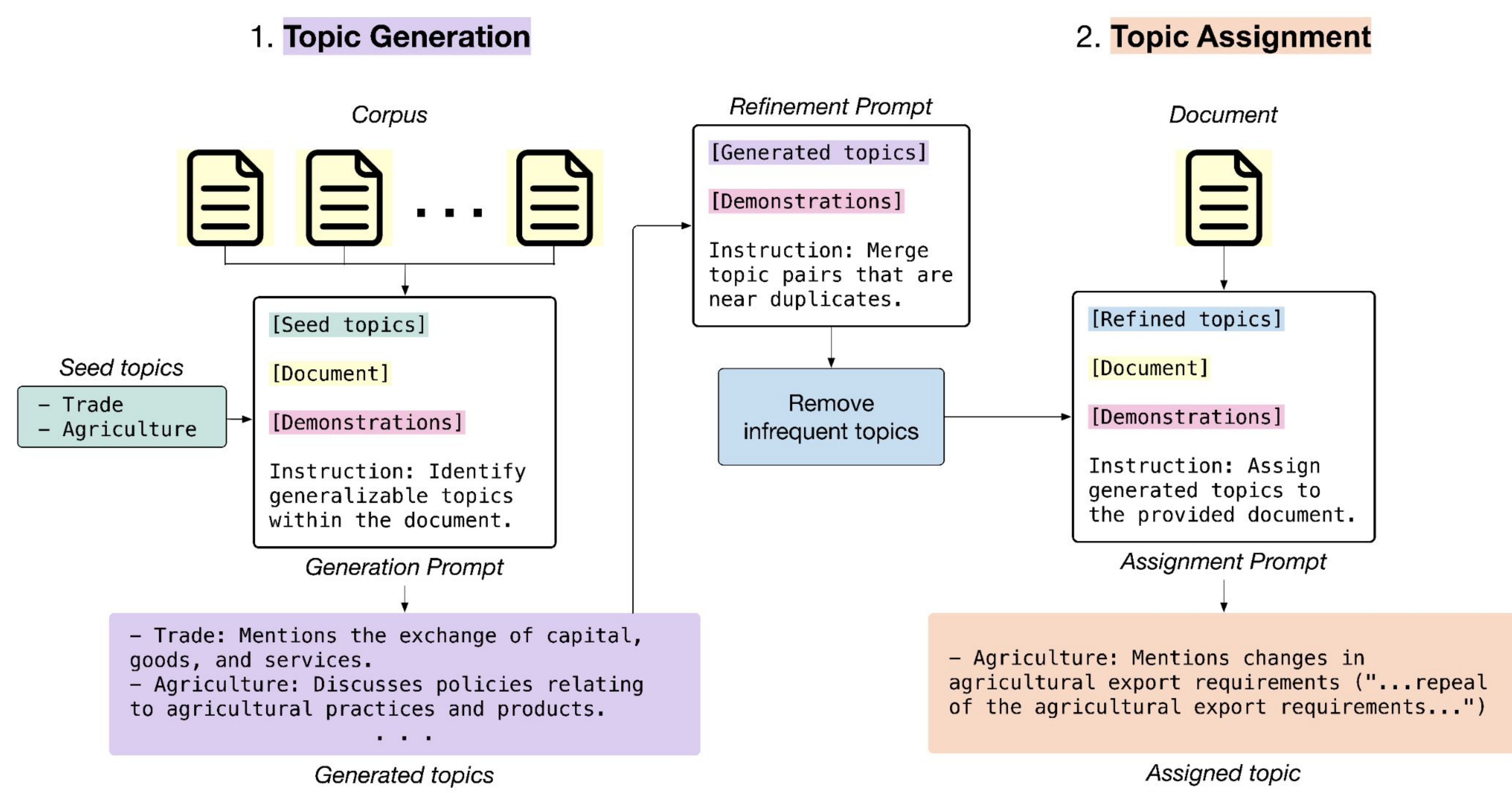
1. TopicGPT produces **interpretable** topics that consist of natural language labels and descriptions.

Wikipedia article

`The Grant Park Music Festival (formerly Grant Park Concerts ) is an annual ten-week classical music concert series held in Chicago, Illinois, USA. It features the Grant Park Symphony Orchestra and Grant Park Chorus along with featured guest performers and conductors….`

- **TopicGPT:** Music & Performing Art (Discuss creation, production, and performance of music, as well as related arts and cultural aspect).
- **LDA:** city, building, area, new, park.

2. TopicGPT is **customizable** to fit user needs through seed topic guidance and semantic-based topic refinement.



3. TopicGPT **outperforms** current state of the art in generating topics that are aligned with human-annotated topics.

- Baselines: LDA, BERTopic.
- Default setting: generator = `GPT-4`, assigner = `GPT-3.5-turbo`.
- Dataset: Wikipedia articles (`Wiki`) and Congressional bills summaries (`Bills`)
- Alignment Metrics: Harmonic Mean Purity (P_1), Adjusted Rand Index (ARI), Normalized Mutual Information (NMI).

Dataset	Setting	TopicGPT			LDA			BERTopic		
		P_1	ARI	NMI	P_1	ARI	NMI	P_1	ARI	NMI
`Wiki`	Default setting ($k = 31$)	**0.73**	**0.58**	**0.71**	0.59	0.44	0.65	0.54	0.24	0.50
	Refined topics ($k = 22$)	**0.74**	**0.60**	**0.70**	0.64	0.52	0.67	0.58	0.28	0.50
`Bills`	Default setting ($k = 79$)	**0.57**	**0.42**	**0.52**	0.39	0.21	0.47	0.42	0.10	0.40
	Refined topics ($k = 24$)	**0.57**	**0.40**	**0.49**	0.52	0.32	0.46	0.39	0.12	0.34
TopicGPT stability ablations, baselines controlled to have the same number of topics (k).										
`Bills`	Different generation sample ($k = 73$)	**0.57**	**0.40**	**0.51**	0.41	0.23	0.47	0.38	0.08	0.38
	Out-of-domain prompts ($k = 147$)	**0.55**	**0.39**	**0.51**	0.31	0.14	0.47	0.35	0.07	0.41
	Additional seed topics ($k = 123$)	**0.50**	**0.33**	**0.49**	0.33	0.15	0.46	0.36	0.07	0.40
	Shuffled generation sample ($k = 118$)	**0.55**	**0.40**	**0.52**	0.33	0.16	0.47	0.36	0.08	0.40
	Assigning with Mistral ($k = 79$)	**0.51**	**0.37**	**0.46**	0.39	0.21	0.47	0.42	0.10	0.40

4. TopicGPT topics are **semantically close** to ground truth.

- Manual matching between ground truth and TopicGPT & LDA labels.
- Misaligned topics are categorized as (1) out-of-scope, (2) missing, or (3) repeated.

Dataset	Setting	Out-of-scope	Missing	Repeated	Total
`Wiki`	LDA ($k = 31$)	46.3	4.3	11.9	62.4
	Unrefined ($k = 31$)	38.7	**0.0**	1.1	39.8
	Refined ($k = 22$)	**30.3**	**0.0**	**0.0**	**30.3**
`Bills`	LDA ($k = 79$)	56.1	2.1	22.0	80.2
	Unrefined ($k = 79$)	65.0	**1.3**	3.8	70.1
	Refined ($k = 24$)	**27.8**	4.2	**0.0**	**31.9**

5. TopicGPT can also be extended to a **hierarchical** setting.

- Generated topics are treated as the top-level topics and LLMs are prompted to generate subtopics.