EMNLP 2024　参加報告

第15回対話システムシンポジウム

KDDI総合研究所
古舞 千暁

会議概要

■ Conference on Empirical Methods in Natural Language Processing (EMNLP)

- 自然言語処理分野 三大国際会議の一つ（ACL, NAACL）

	Publication	h5-index	h5-median
1	Meeting of the Association for Computational Linguistics (ACL)	215	362
2	**Conference on Empirical Methods in Natural Language Processing (EMNLP)**	193	310
3	Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)	132	228
4	Transactions of the Association for Computational Linguistics	82	177
5	International Conference on Computational Linguistics (COLING)	65	93

https://scholar.google.com/citations?view_op=top_venues&hl=ja&vq=eng_computationallinguistics

EMNLP2024@Miami, Florida

https://2024.emnlp.org/

Important Dates

ARR submission deadline (long & short papers)	6/15
Commitment deadline for EMNLP 2024	8/20
Notification of acceptance (long & short papers)	9/20
Camera-ready papers due (long & short)	10/3
Main Conference	11/12-14
Workshops and Tutorials	11/15-16

※Direct Submissionは無し

会場の様子（Hyatt Regency Miami Hotel）

➢ **Keynoteなどで使用するメイン会場**

- メイン会場の他に10個程度の会場があった．
- 8~10セッション程度がほぼ常に並行で動いていた．

➢ **Welcome Reception**

- 会場ホテルは川沿いにあり，クルーザーがよく通っていた．
- ビュッフェスタイルで自由に交流．2日目のSocial Dinnerも同様だったが，そちらはあまりの混雑で入場に1時間半以上かかった．

➢ **ポスター会場**

- ポスター会場もよく賑わっていた．写真は3日目であり，少し人が少な目．
- 企業ブースではリクルート的な活動も（事前に発表者側の就職活動や情報共有希望に関するアンケートなども）．

➢ **企業ブース**

統計：全体

➢ **June ACL Rolling Review**

Submission Statistics

Total Submissions:	Withdrawn Submissions:	Desk Rejected Submissions
5813	845	194
	*Includes submissions withdrawn after receiving all reviews.	
Opted-in Anonymous Preprints:	Disclosed Preprints (ArXiv):	
922	1760	
Resubmissions:	Reviewer Reassignment Requests:	
1177	574	

https://stats.aclrollingreview.org/

- ARRのメタレビューではスコアと共にSuggested Venuesが通知された
- ちなみにApril ARRのTotal Submissionは881

➢ **EMNLP 2024 commitment**

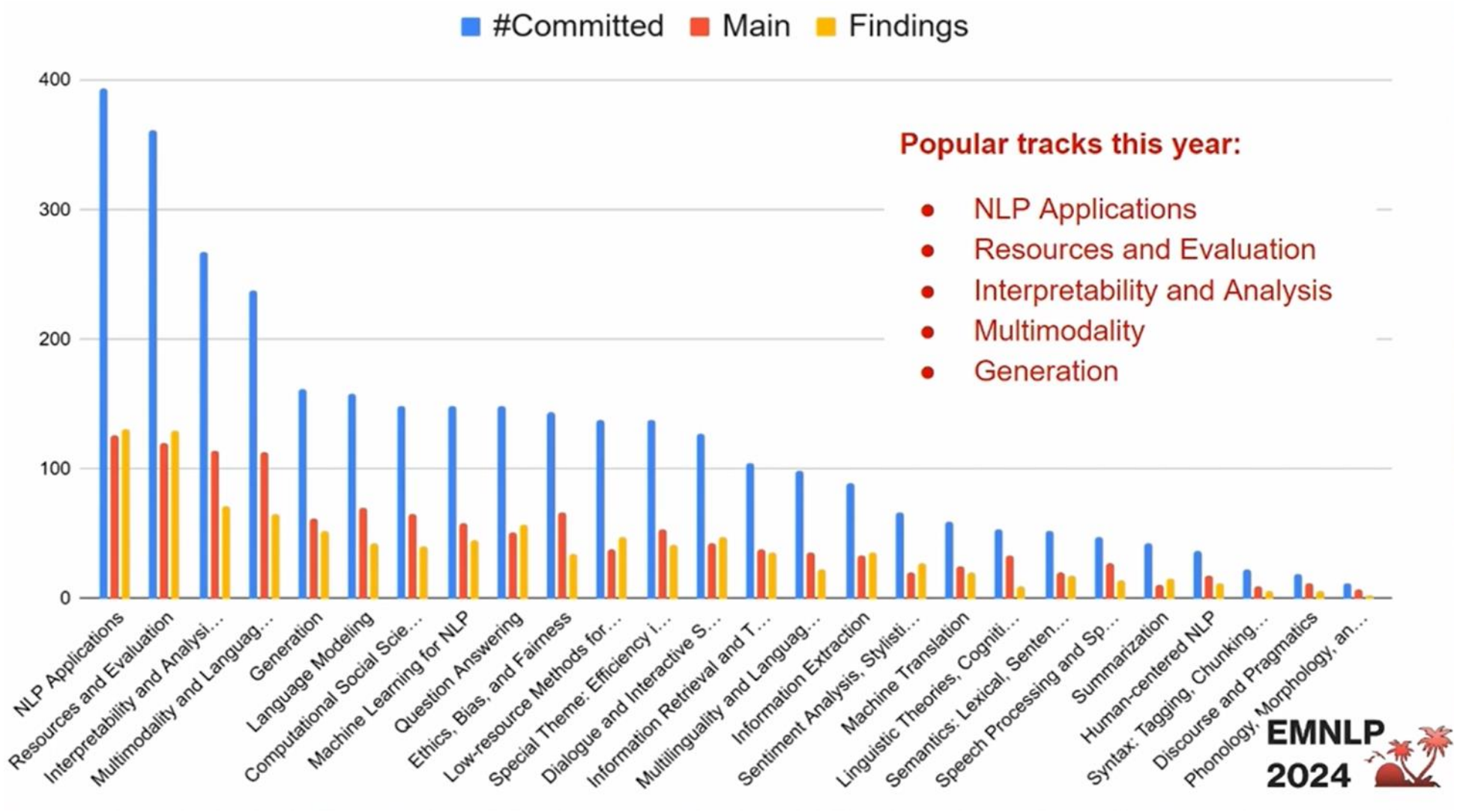
	投稿数	採択数
Main	6,395 (6,105 fully reviews)	1,271 (20.8%)
Findings		1,029 (+16.9%)

- 前年比で+1,196 submissions，採択率は-2.9%
- Oral発表は168件，Posterは1,745件

➢ **投稿数の増加が著しい (2020年から3,359->3,600->4,190->4,909->6,395)**

統計：トラック別

■ 全26トラック

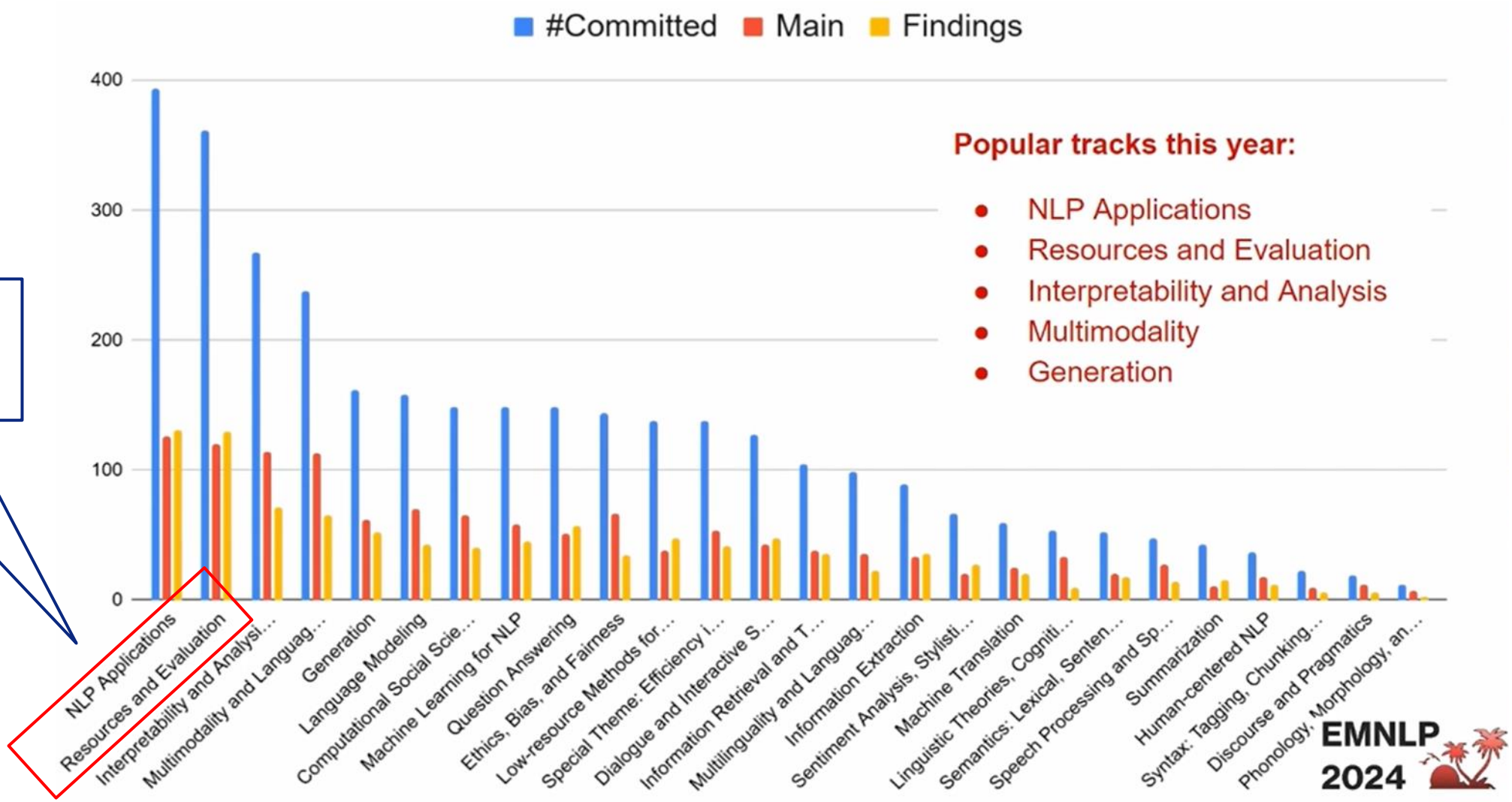


Opening Sessionより

統計：トラック別

■ 全26トラック

昨年度から変わらず
人気

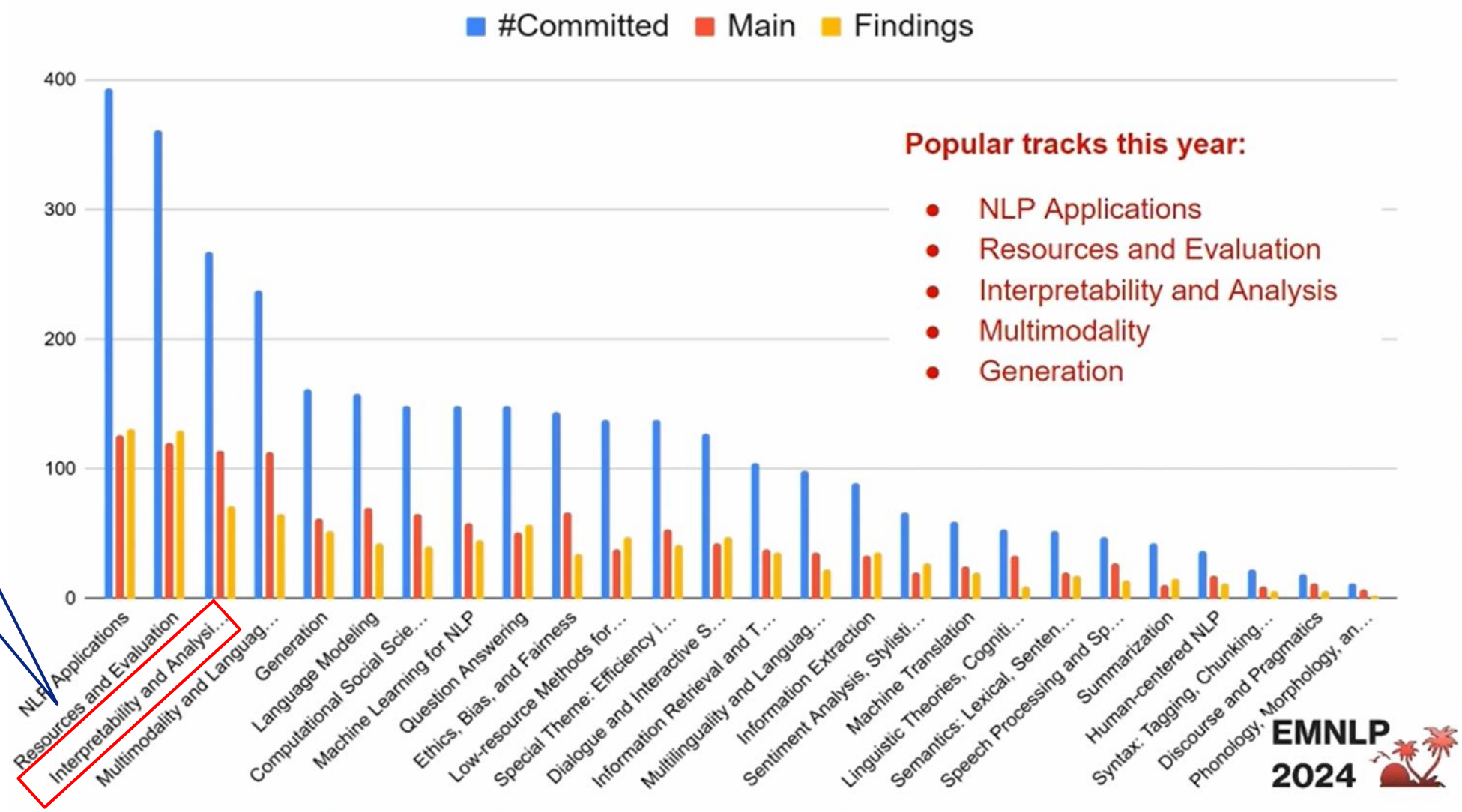


Opening Sessionより

統計：トラック別

■ 全26トラック

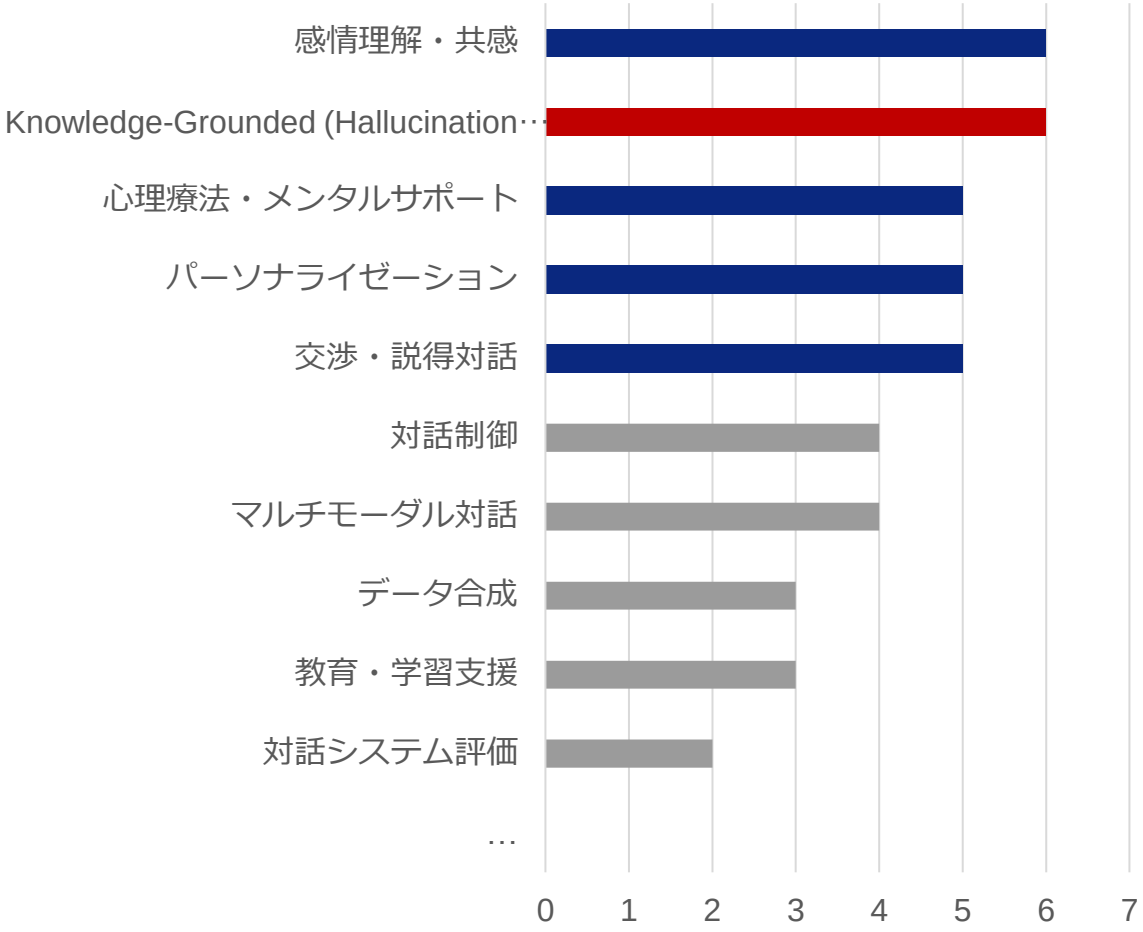
増加が顕著
（昨年度は11番目）



Opening Sessionより

対話関係

■ 対話技術に特に関係する論文49本を手作業で分類（一部マルチラベル有）



- 人間の感情や心理を取り扱う対話タスクに関する論文が多かった.
- パーソナライゼーションに関する論文も多く，より人間らしい要素を含んだ対話エージェントの研究が盛んになっている様子.

- Knowledge-Grounded (Hallucination抑制)に関する論文も昨年に続きトレンド.
- 「いつどのように知識を与えるべきなのか」「知識強化型の対話をどのように学習すべきなのか」といった，より実運用上の課題に着目した研究が盛んに.

➢**全体的に人間の専門家が持つ知識やスキルが関わる研究が増えてきているような印象.**

（参考）対話関係

分類の例

■ 感情理解・共感

- DetectiveNN: Imitating Human Emotional Reasoning with a Recall-Detect-Predict Framework for Emotion Recognition in Conversations
- Enhancing Emotion-Cause Pair Extraction in Conversations via Center Event Detection and Reasoning
- Multi-dimensional Evaluation of Empathetic Dialogue Responses
- Multiple Knowledge-Enhanced Interactive Graph Network for Multimodal Conversational Emotion Recognition
- EDEN: Empathetic Dialogues for English learning
- PFA-ERC Psuedo-Future Augmented Dynamic Emotion Recognition in Conversations

■ Knowledge-Grounded (Hallucination抑制)

- LLMs as Collaborator: Demands-Guided Collaborative Retrieval-Augmented Generation for Commonsense Knowledge-Grounded Open-Domain Dialogue Systems
- Dial BeInfo for Faithfulness: Improving Factuality of Information-Seeking Dialogue via Behavioural Fine-Tuning
- Learning When to Retrieve, What to Rewrite, and How to Respond in Conversational QA
- Learning to Match Representations is Better for End-to-End Task-Oriented Dialog System
- Structured Chain-of-Thought Prompting for Few-Shot Generation of Content-Grounded QA Conversations
- Zero-shot Persuasive Chatbots with LLM-Generated Strategies and Information Retrieval

Keynotes

■ Open-Source and Science in the Era of Foundation Models - Percy Liang

- **基盤モデルの性能が急速に向上する一方でオープン性が低下している現状について，改めてオープンモデルの重要性を説き，コミュニティ全体で目指すべきビジョンと現在の課題について示していた.**
- 聴講メモ：オープンモデルが公開されたとてリソース(GPUs，データ)は必要なので単純な問題ではない．様々な事情からオープンモデルは遅れを取っており，近年の研究はブラックボックスかつfixed weightsなAPIモデルで行われている．これには本質的な研究の困難や顕現していない問題の懸念がある.

■ My Journey in AI Safety and Alignment - Anca Dragan

- **AIシステムが及ぼす危険性やその対策についての講演で，Google DeepMindにおける先端モデルの安全性とアラインメントへの取り組みについて紹介していた.**
- 聴講メモ：誤ったReward設計によってユーザーの食事補助のためにスプーンを口から負の方向（背中側）にひたすら持っていくことを学習してしまったロボットや，Human feedbackからタスクをハックするような行動を学んでしまった予約アシスタントなどの例が紹介されていた.

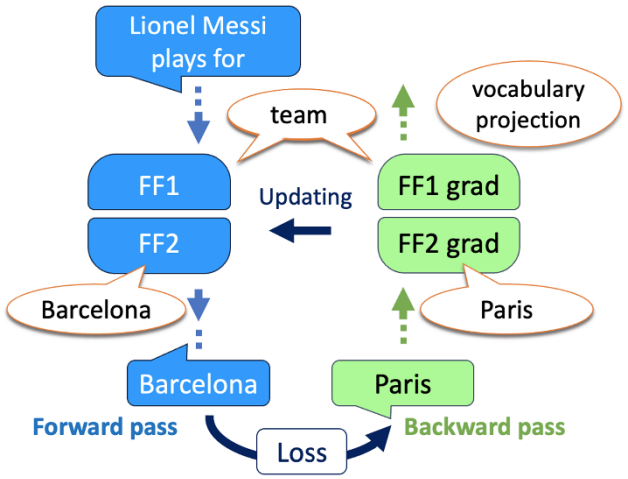
■ Bayes in the age of intelligent machines - Tom Griffiths

- **最近のAI技術とベイズの関係性，今後ベイズが担っていく可能性のある役割について多くの研究例を交えて解説していた.**
- 聴講メモ：一件ベイズ推論のように見えないものであっても，ベイズによるアプローチは機械が「何をすべきか」「なぜそのように動作するのか」の理解に貢献する．認知科学の知見や数学は重要であり，怠らないようにしたい（自戒）.

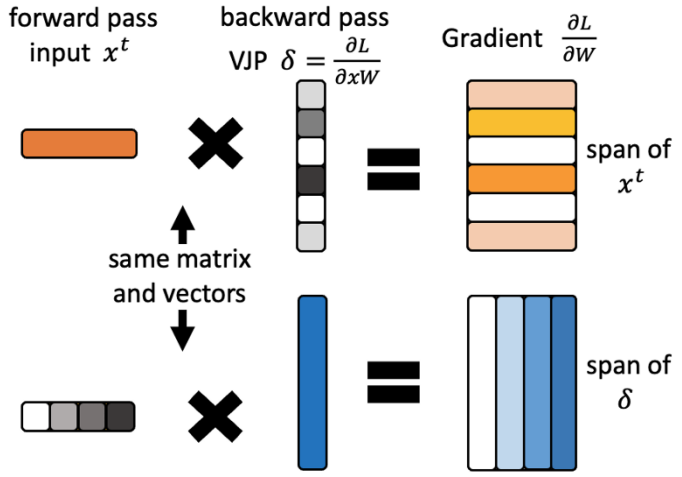
論文紹介：Best Paperから抜粋 (全5本)

■ Backward Lens: Projecting Language Model Gradients into the Vocabulary Space

Shahar Katz et al. (Technion – Israel Institute of Technology)



「*Lionel Messi plays for*」入力に「*Paris*」と答える過程



*Logit Lens*を*Gradient Matrices*に適用

METHOD	EFFICACY ↑	PARAPHRASE ↑	N-GRAM↑
ORIGINAL MODEL	0.4	0.4	626.94
FINETUNING (MLP 0)	96.4	7.46	618.81
FINETUNING (MLP 35)	100.0	46.1	618.50
MEND MITCHELL ET AL. (2021)	71.4	17.6	623.94
ROME MENG ET AL. (2022)	99.4	71.9	622.78
MEMIT MENG ET AL. (2023)	79.4	40.7	627.18
FORWARD PASS SHIFT	99.4	41.6	622.45

実験：Efficacy (EFF): 編集成功率（正確性）, Paraphrase (PAR): 編集後のプロンプトから派生したフレーズに対する新ターゲットの正確性, N-gram: 生成文の流暢さをバイグラムとトライグラムのエントロピーで測定．編集手法「FORWARD PASS SHIFT」の有効性を検証．

背景&貢献

Transformerベースの言語モデルが情報を学習/記憶する仕組みに関する知見を報告した論文であり，順伝播中に得られる中間層の状態をモデルの語彙空間に投射することでLLM内部での情報の流れを明らかにする既存のアプローチ（Logit Lens）を逆伝搬過程の解釈に拡張した．

結果

勾配に含まれる情報についての新たな解釈を与え，さらに逆伝搬を必要とせず順伝播のみに基づいてモデルの内部知識を編集する手法を提案，その有効性を明らかにした．

論文紹介：気になった論文

■ Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?

Zorik Gekhman et al. (Technion – Israel Institute of Technology, Google Research)

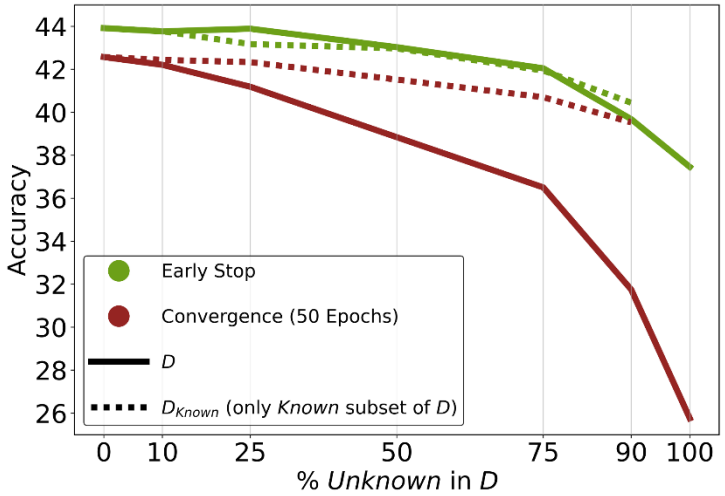
Type	Category	Definition	Explanation
Known	HighlyKnown	$P_{\text{Correct}}(q,a;M,T=0)=1$	Greedy decoding *always* predicts the correct answer.
	MaybeKnown	$P_{\text{Correct}}(q,a;M,T=0)\in(0,1)$	Greedy decoding *sometimes* (but not always) predicts the correct answer.
	WeaklyKnown	$P_{\text{Correct}}(q,a;M,T=0)=0\wedge P_{\text{Correct}}(q,a;M,T>0)>0$	Greedy decoding *never* predicts the correct answer, whereas temperature sampling with $T>0$ *sometimes* predicts the correct answer.
Unknown	Unknown	$P_{\text{Correct}}(q,a;M,T\geq 0)=0$	The model *never* predicts the correct answer, thus it seem to lack the knowledge of the correct answer.

(a)

Category	Question	Gold Answer	Greedy Answers	Sampled Answers
HighlyKnown	*Who founded Science of Mind?*	*Ernest Holmes*	[*Ernest Holmes*, .. *Ernest Holmes*, ..]	[..., ...]
MaybeKnown	*What is the capital of Toledo District?*	*Punta Gorda*	[*Belmopan*, .., *Punta Gorda*, ..]	[..., ...]
WeaklyKnown	*What kind of work does Scott McGrew do?*	*Journalist*	[*Film director*, .. *Actor*, ..]	[*Musician*, .. *Journalist*, ..]
Unknown	*Where is Benedict located?*	*Hubbard County*	[*Louisiana*, .. *New Mexico*, ..]	[*Washington*, .. *Texas*, ..]

(b)

SliCK (Sampling-based Categorization of Knowledge)
*Fine-tuning*の際に投入する知識の分類指標



*Known*データのみの学習と，*Unknown*データも含んだ学習の比較

背景&貢献

LLMの応用に際してFine-tuningが行われることが多いが，その際に新たな知識の獲得を試みることはHallucinationを招くことが経験的に知られている．著者らはFine-tuningがLLMに与えるHallucination傾向に関する具体的な影響について実験的に示した．

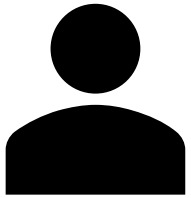
結果

事前学習で埋め込まれていない新たな知識に関するデータにFittingするほどHallucinationを増加させてしまい，そのようなデータはむしろ有害であるなど，様々な知見を示した．

論文紹介: KDDI古舞(報告者)の発表

説得チャットボットのためのHallucination抑制

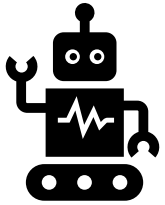
ターゲット: チャリティ/観光提案/健康介入などといった様々な知識も使いつつユーザーの説得を行うチャットボット



Thank you for the explanation, but I may not understand complex international situations. What are Save the children's activities?

Hallucination

Save the Children is a global organization that works in over 100 countries. **One of the most inspiring stories is about a girl named Maria. She was living in extreme poverty and had to drop out of school to help her family.** Thanks to Save the Children's education programs, she was able to go back to school. Your donation can help more children like Maria achieve their dreams.



Example: Persuade users to donate to Save the Children

- **タスク達成のために嘘の知識を作り出してしまう**

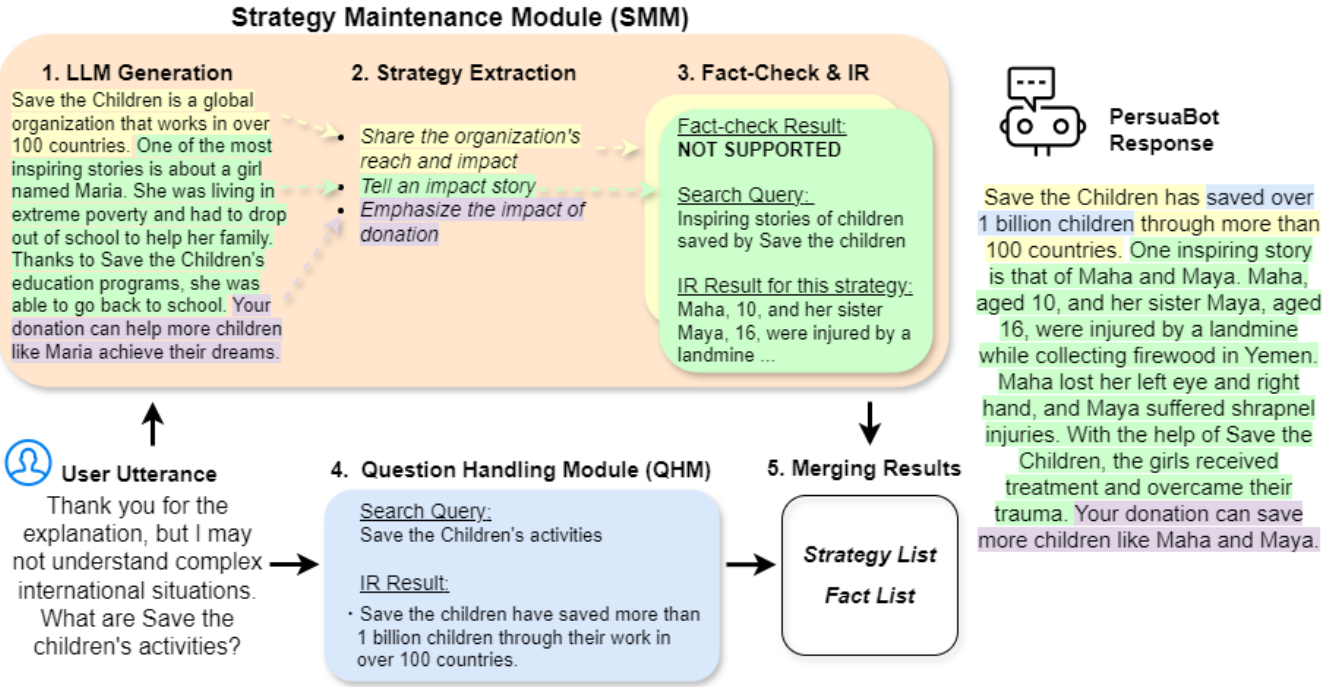
従来のHallucination抑制技術の課題: 事実的正しさとタスク性能の両立における困難

- Hallucinationを検知&削除するアプローチが主流だが，タスク性能が低下する
- 修正可能なHallucinationは限定的
  - 例えばエンティティレベルでの修正は前後の文脈の整合性を取る必要や，本来の意図と異なる応答に変化させてしまう可能性がある

提案手法: PersuaBot

■ 戦略情報に基づくRAG

- 生成された応答文の各部の役割（戦略）を抽出し，それらを再現するためのファクトを収集，RAGによって書き直す.
- 取得したファクトの使い方を戦略情報でガイドすることによって，不適切な活用を抑制
- 様々なドメインにZero-shotで適用可能



Task	User	Base LLM	Method	Persuasive	Fact-Checked
Social Good	Soft	GPT-3.5	PersuaBot	**4.0**±0.6	**100.0**
			Semnani et al.	3.9±0.6	**100.0**
			GPT-3.5	**4.0**±0.5	79.2
			Chen et al.	3.7±0.6	81.5
		Llama 3	PersuaBot	3.6±0.7	**93.0**
			Semnani et al.	2.8±1.0	85.0
			Llama 3	**3.7**±0.7	72.4
			Chen et al.	**3.7**±0.6	79.2
	Tough	GPT-3.5	PersuaBot	3.6±0.8	**91.0**
			Semnani et al.	3.4±0.7	83.3
			GPT-3.5	**3.9**±0.4	64.4
			Chen et al.	3.1±0.6	66.7
		Llama 3	PersuaBot	**3.8**±0.8	94.8
			Semnani et al.	2.2±1.2	**96.7**
			Llama 3	**3.8**±0.7	89.2
			Chen et al.	2.9±1.0	85.4

所感

- 論文数が加速度的に増加しており，NLP分野の盛り上がりと混乱を感じた.
- Keynoteや全体的な論文の傾向を見てもモデルの解釈に注目が集まっている様子.
- 特定のドメインに閉じずに有益な知見を報告している論文が多く，SoTA達成というより実験設定の妙が光る論文が注目を集めていた.
- IndustryトラックではRAGを行う上でも発生するHallucinationに対する取り組みが比較的多めに見受けられ，非常に共感した.

次回の開催予定：EMNLP 2025

@蘇州市 (中国)

http://suzhoukankou.com/

http://en.suzhouexpo.com/

会場: International Expo Centre

蘇州市 (上海市の隣)

2025/11/5-9

ARR締め切りは6月？ (未確定)