آزمایشگاه پردازش زبان طبیعی
دانشگاه صنعتی امیرکبیر

غنی سازی سامانه‌های گفتگو با مالتی مدیا

Enriching dialogue systems with multimedia

ارائه دهنده:

امیرحسین کریمی

اساتید راهنما:

دکتر اکبری

دکتر محدث

آبان ۱۴۰۰

۰۱ **مقدمه**

بیان صورت مسئله و
انگیزه پایان نامه

۰۲ **مجموعه داده**

بررسی مجموعه
داده‌های مرتبط

۰۳ **چالش‌ها و ارزیابی و کارهای مرتبط**

چالش‌ها و متریک های
ارزیابی سیستم‌های گفتگو

۰۴ **قدم بعدی**

ایده و روش پیشنهادی برای
حل چالش‌های مسئله

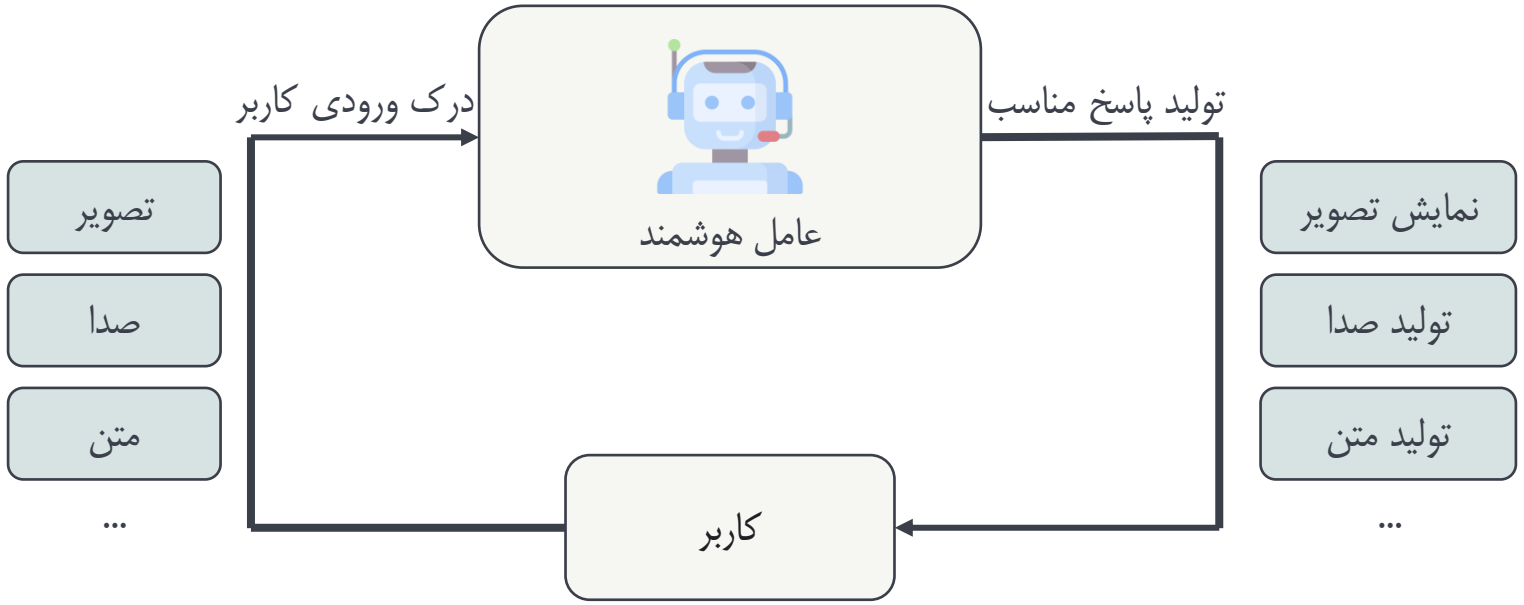
مقدمه

بیان صورت مسئله و انگیزه پایان نامه

مقدمه

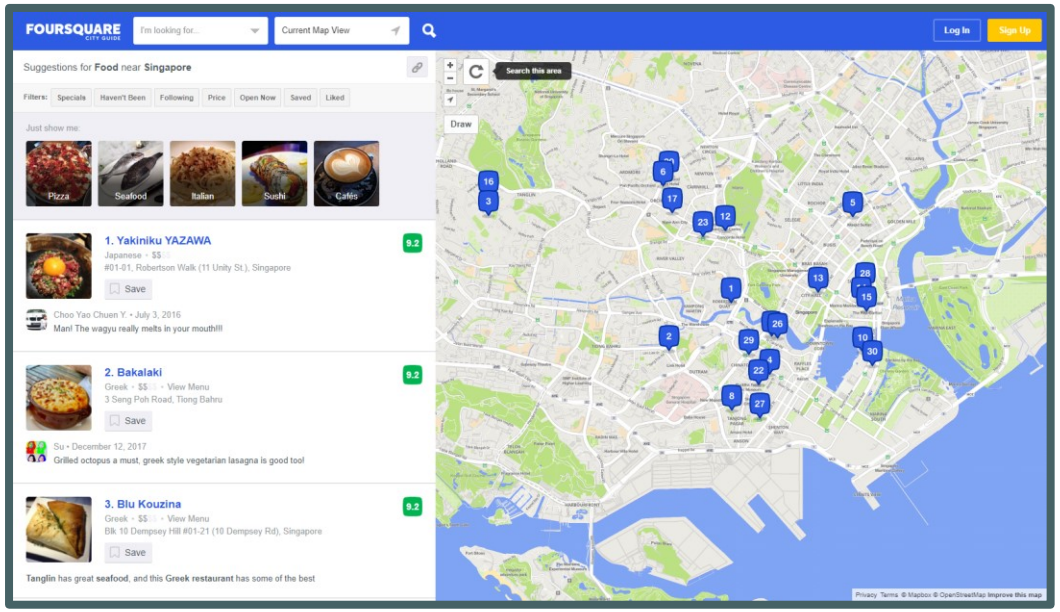
نمایی از یک سیستم چند رسانه (Multimodal System)

هدف نهایی سیستم‌های هوشمند، تعامل با کاربر به شکل چند رسانه است.

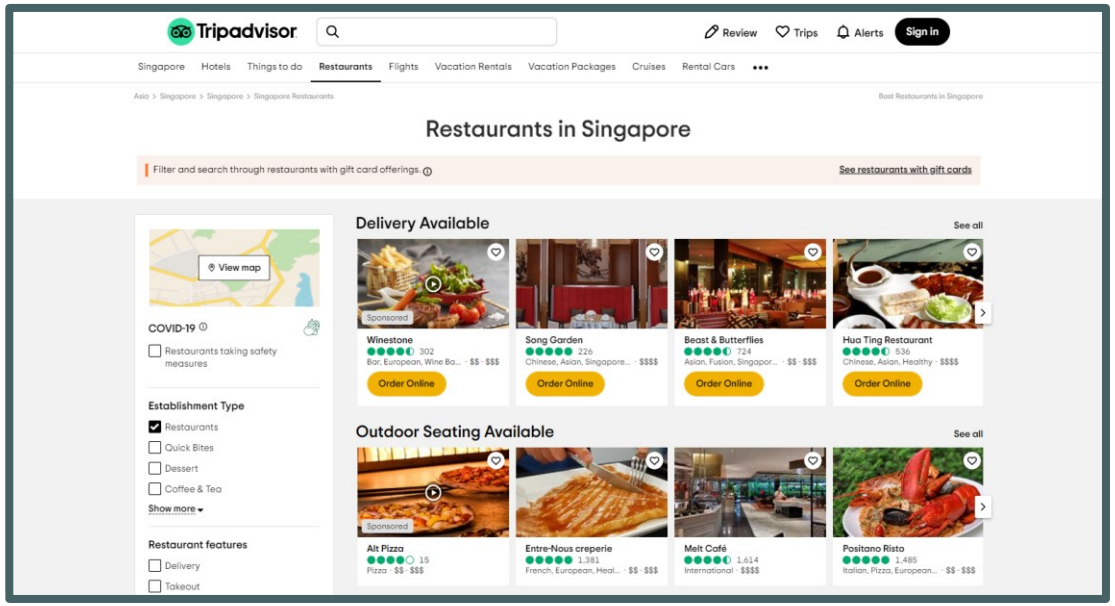


انگیزه پایان‌نامه

جستجوی چندرسانه برای مکان‌های جدید



نمایی از سایت foursquare.com



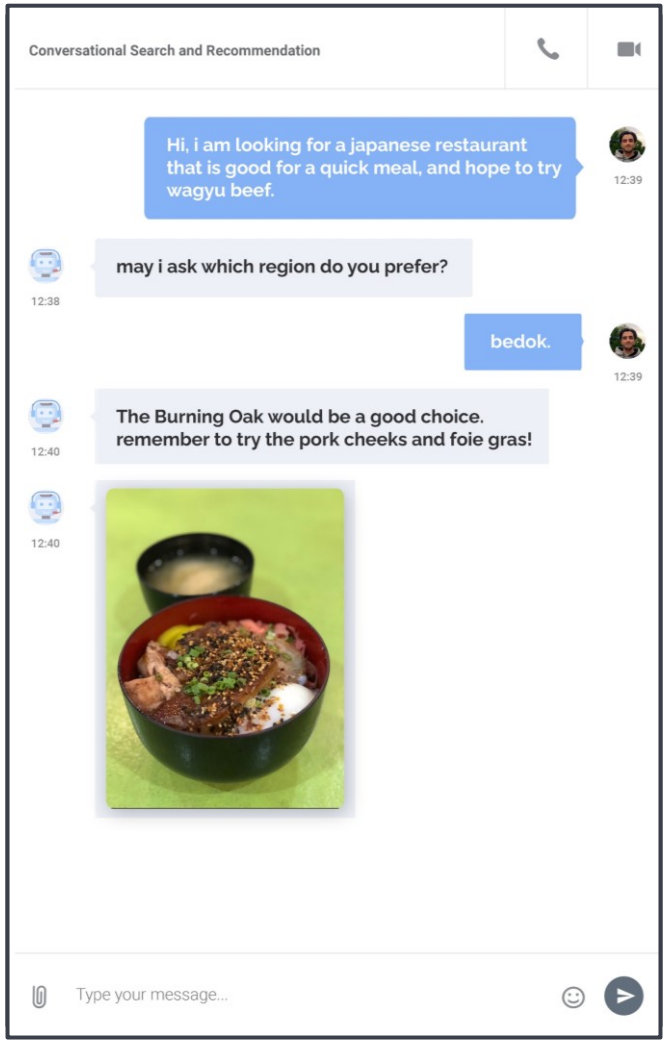
نمایی از سایت tripadvisor.com

انگیزه پایان‌نامه

مثالی از سیستم گفتگوی تک رسانه

سیستم تک حالته متنی، به کاربر کمک کرد که رستوران مناسب را پیدا کند و سیستم غذای جدیدی به کاربر پیشنهاد داد،

اما کاربر نمی‌داند با چه نوع غذایی رو به رو خواهد شد.



تصویر از نویسنده - ساخته شده بر اساس مجموعه داده MMConv

همان گفتگو اگر به شکل چند رسانه باشد

سیستم چندرسانه به کاربر کمک می‌کند مکان‌ها یا غذاهای جدیدی پیدا کند و **با دیدن تصاویر آگاهانه‌تر تصمیم بگیرد و رضایت بیشتری از این تعامل با سیستم داشته باشد.**

مجموعه داده

بررسی مجموعه داده‌های مرتبط

مجموعه داده

مقایسه مجموعه‌داده‌های مختلف در سیستم‌های توصیه‌گر و سیستم‌های گفتگو

	Datasets	# Dialogues	# Utters	Types	Domains	User Data	Modality	State Label
توصیه‌گر مکالمه‌ای	Facebook Rec [8]	1M	6M	Conv. Rec.	Movie	×	Text	×
	REDIAL [17]	10K	163K	Conv. Rec.	Movie	×	Text	×
	TG-ReDial [44]	10K	129K	Conv. Rec.	Movie	√	Text	×
	OpenDialKG [23]	15K	143K	Conv. Rec.	Movie, book	×	Text	×
	DuRecDial [21]	10K	156K	Conv. Rec.	Movie, music, news etc.	√	Text	×
	MGConvRex [40]	7K	73K	Conv. Rec.	Restaurant	√	Text	√
جستجوگر مکالمه‌ای	WOZ 2.0[25]	1.2K	12K	Conv. Search	Restaurant	×	Text	√
	DSTC2 [38]	1.6K	23K	Conv. Search	Restaurant	×	Text	√
	FRAMES [9]	1.3K	20K	Conv. Search	Flight, hotel, budget	×	Text	√
	KVRET [10]	3K	15K	Conv. Search	In-car assistant	×	Text	×
	MultiWOZ [3]	8K	115K	Conv. Search	Hotel, restaurant etc.	×	Text	√
چندرسانه	VisDial [5]	123K	2.4M	Image-based QAs	Concepts in image	×	Multi.	×
	GuessWhat [6]	155K	1.6M	Image-based QAs	Concepts in image	×	Multi.	×
	IGC [24]	4K	25K	Image-based QAs	Concepts in image	×	Multi.	×
	MMD [29]	150K	6M	Fashion Search	Fashion	×	Multi.	×
	MMConv	5.1K	39.7K	Conv. Search	5 domains in travel	√	Multi.	√

Liao et al. “MMConv: An Environment for Multimodal Conversational Search across Multiple Domains”. (2021)

مجموعه داده

مجموعه داده چندرسانه MMConv

نمایی از نحوه جمع‌آوری مجموعه داده چندرسانه MMConv به شکل انسان–با–انسان. کاربر باید از پروفایل کاربر تعیین شده پیروی کند و عامل باید با استفاده از متن و تصویر و جستجو در پایگاه داده، جواب مناسبی به کاربر دهد.

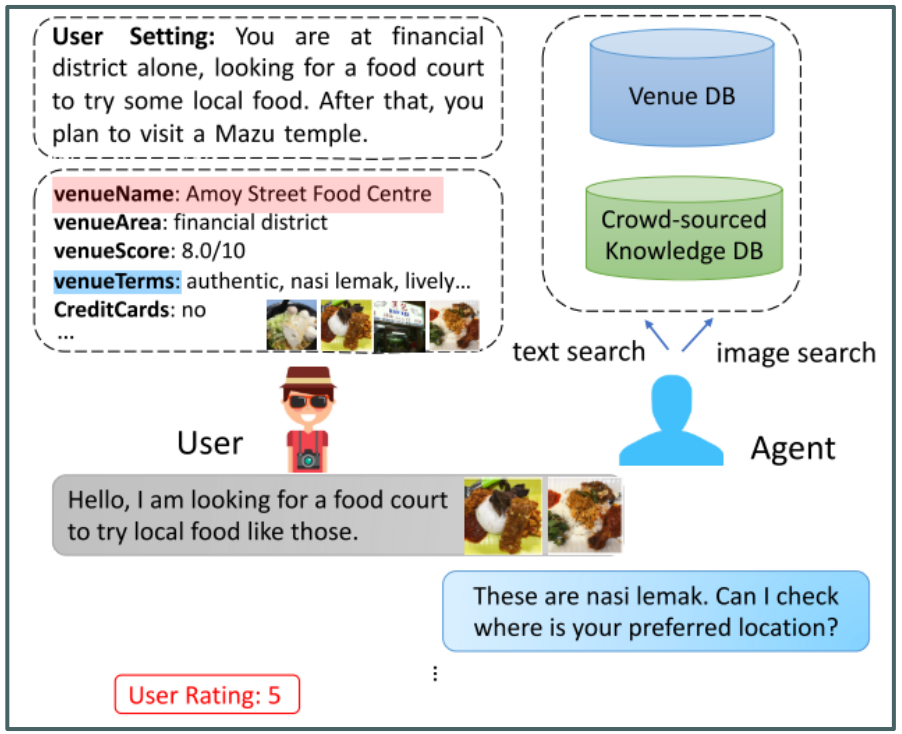


Table 2: Venue distribution over different domains.

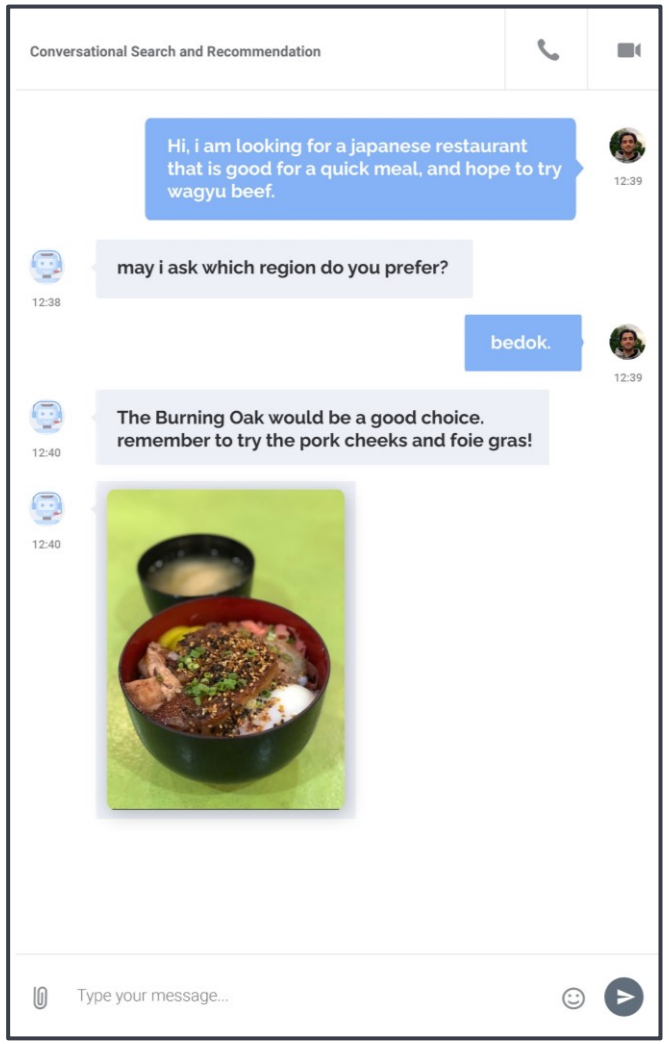
Domain	food	hotel	nightlife	mall	sightseeing
# Venues	1,162	83	128	96	302
# Reviews	27,303	2,661	3,064	3,803	6,019
# Images	67,058	2,268	12,692	851	31,084

Table 3: Statistics of different conversation scenarios.

Scenarios	# conversations
Search venue by image	104
Recognize concept by image	214
Find venue by preferences	4,924
Cross-domain venue recommendation	3,007
Subsequent venue substitution	683
Venue comparison	182
Find specific shop in mall	53

Liao et al. “MMConv: An Environment for Multimodal Conversational Search across Multiple Domains”. (2021)

غنی سازی گفتگو



تصویر از نویسنده - ساخته شده بر اساس مجموعه داده MMConv

دو بخش اصلی برای غنی سازی گفتگو:

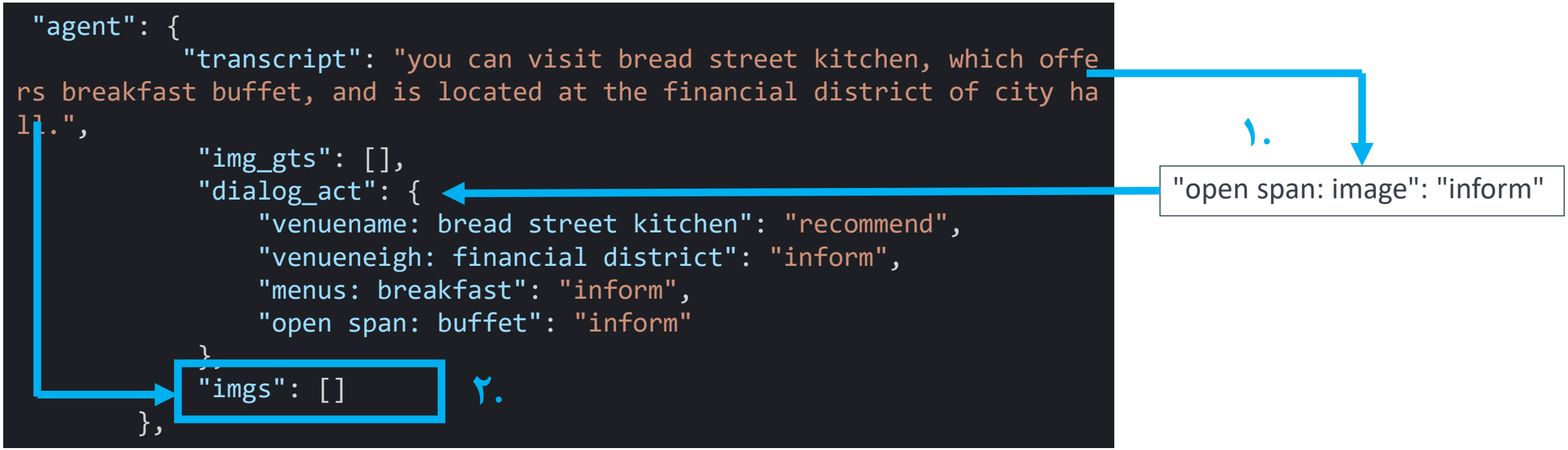
- ❑ تعیین زمان مناسب در بین گفتگو برای نمایش تصویر
- ❑ پیدا کردن تصویر مرتبط به متن گفتگو (برای مثال از پایگاه داده)

غنی سازی گفتگو

دو بخش اصلی برای غنی سازی گفتگو:

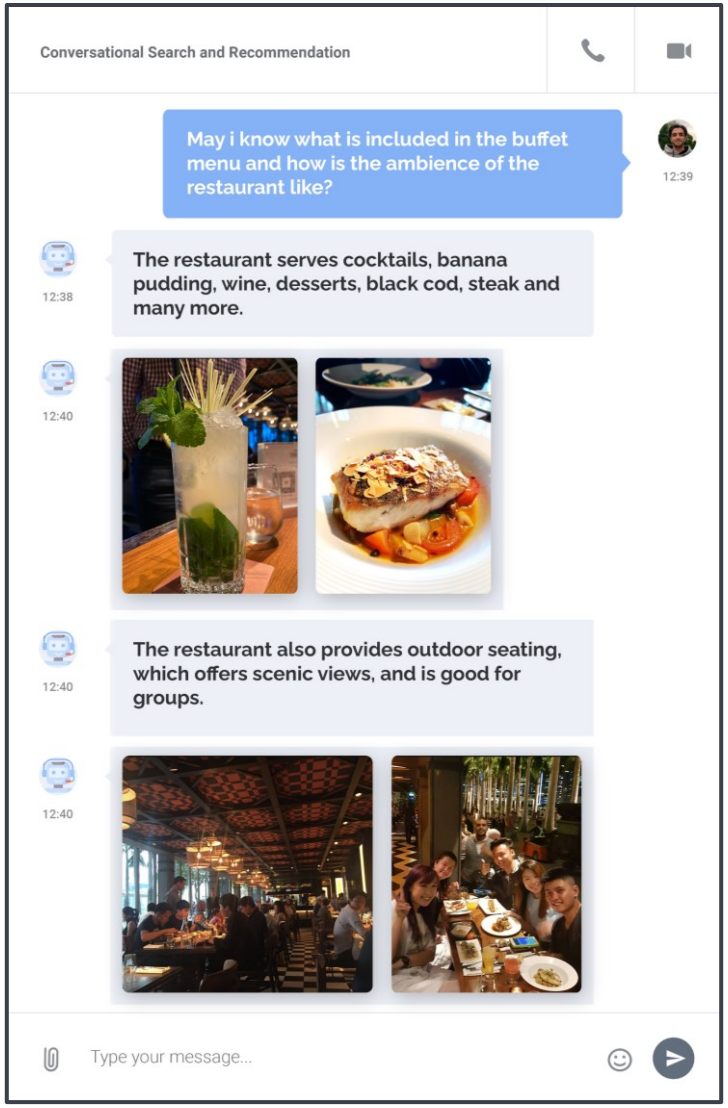
۱. تعیین زمان مناسب در بین گفتگو برای نمایش تصویر (Action prediction)

۲. پیدا کردن تصویر مرتبط به متن گفتگو از پایگاه داده (Image Retrieval)



نمونه هدف نهایی پایان نامه

نمونه‌ای از یک سیستم توصـیه‌گر گفتگو محور با مالتی مدیا:



تصویر از نویسنده - ساخته شده بر اساس مجموعه داده MMConv

چالش‌ها

برخی از مهم‌ترین چالش‌های مجموعه داده MMConv

چالش‌ها – اسلات‌ها

اسلات‌ها را می‌توان به دو دسته کلی تقسیم کرد:

- اسلات‌های قابل درخواست (requestable slots) که یک سری اطلاعات منحصر به فرد مکان هستند *مانند آدرس یا شماره تلفن*
- اسلات‌های قابل اطلاع (informable slots) ویژگی‌هایی که کاربر به سیستم می‌دهد تا گزینه‌ها را محدود کند *مانند محدوده قیمت*

اما کاربران اصطلاحاتی برای مکان‌ها استفاده می‌کنند مانند: family friendly، good for groups، great value و ...

استفاده از این اطلاعات در پیدا کردن مکان‌ها ضروری است اما گنجادن آن‌ها در قالب اسلات دشوار است.

این اطلاعات در MMConv وجود دارد و دسته بندی نشده‌اند و تمام آن‌ها را برچسب open span در مجموعه داده قرار گرفته‌اند.

نتایج ضعیف روش‌های لبه دانش بر روی این مجموعه داده، نشان‌دهنده‌ی نیاز به مدل‌های پیشرفته‌تر است.

Liao et al. “MMConv: An Environment for Multimodal Conversational Search across Multiple Domains”. (2021)

چالش‌ها – تصاویر

برخی از موارد استفاده تصاویر در مجموعه داده:

- نمایش غذا و محتویات آن
- پیدا کردن مکان‌ها و غذاهای مرتبط به تصویر کاربر
- توصیف فضا و اتمسفر مکان‌ها

در مجموعه‌داده‌هایی مانند MultiWOZ یا DSTC2 اطلاعات و حاشیه‌نویسی گفتگو وجود دارد اما فقط متنی و تک حالته هستند.

در مجموعه‌داده‌هایی مانند VisDial یا MMD تصاویر و ویژگی‌های آنها وجود دارد اما اطلاعات و حاشیه‌نویسی برای گفتگو وجود ندارد.

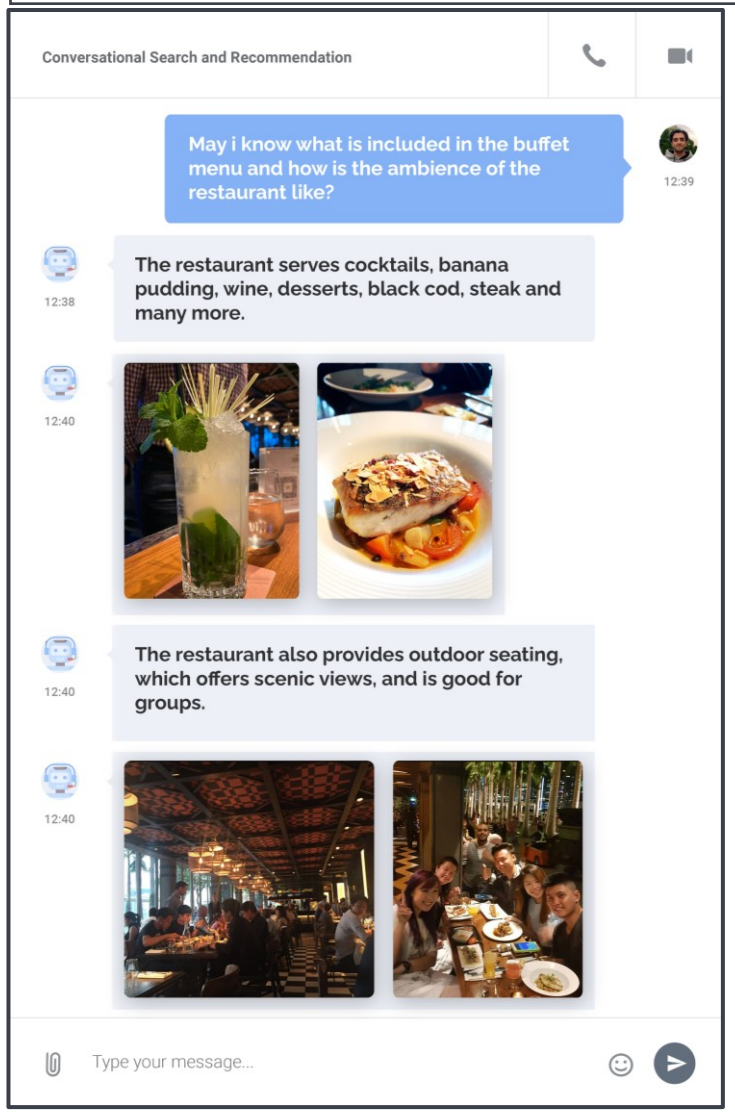
روش‌های موجود برای Response generation یا Dialogue state tracking برای مدریت تصاویر در گفتگو با مشکل مواجه هستند.

Liao et al. “MMConv: An Environment for Multimodal Conversational Search across Multiple Domains”. (2021)

وظایف سامانه و نحوه ارزیابی

متریک های ارزیابی سیستم‌های گفتگو

وظایف سامانه



تصویر از نویسنده - ساخته شده بر اساس مجموعه داده MMConv

۱. انتخاب عمل و ارسال درست درخواست به API

مشخص کردن (مانند مشخص کردن قیمت هتل یا رستوران)

جستجو کردن (مانند انتخاب تصویر مرتبط)

۲. تولید پاسخ مناسب

۱) مانند مدل زبانی

۲) مانند بازیابی اطلاعات

۳. دنبال کردن حالات گفتگو (DST)

ارزیابی وظایف سامانه

ارزیابی:

۱. انتخاب عمل و ارسال درست درخواست به API

- ❑ Accuracy
- ❑ Perplexity

۲. تولید پاسخ مناسب

۱) مانند مدل زبانی

۲) مانند بازیابی اطلاعات

- ❑ BLEU
- ❑ Recall@k

۳. دنبال کردن حالات گفتگو (DST)

- ❑ F1

دیگر ارزیابی‌ها

Inform rate

- تعیین می‌کند آیا سیستم گفتگو، موجودیت‌های مربوطه را ارائه می‌کند یا خیر (برای مثال نام رستوران).

Success rate

- موفقیت تکلیف را اندازه‌گیری می‌کند، نسبت مکالمه‌هایی که مشکل کاربر را با موفقیت حل کرده است (مانند یافتن مکان مناسب برای کاربر).

Combined

- گاهی امتیاز ترکیبی Budzianowski et al. (2018) نیز با فرمول زیر گزارش می‌شود:

$$\text{Combined} = (\text{Inform} + \text{Success}) \times 0.5 + \text{BLEU}$$

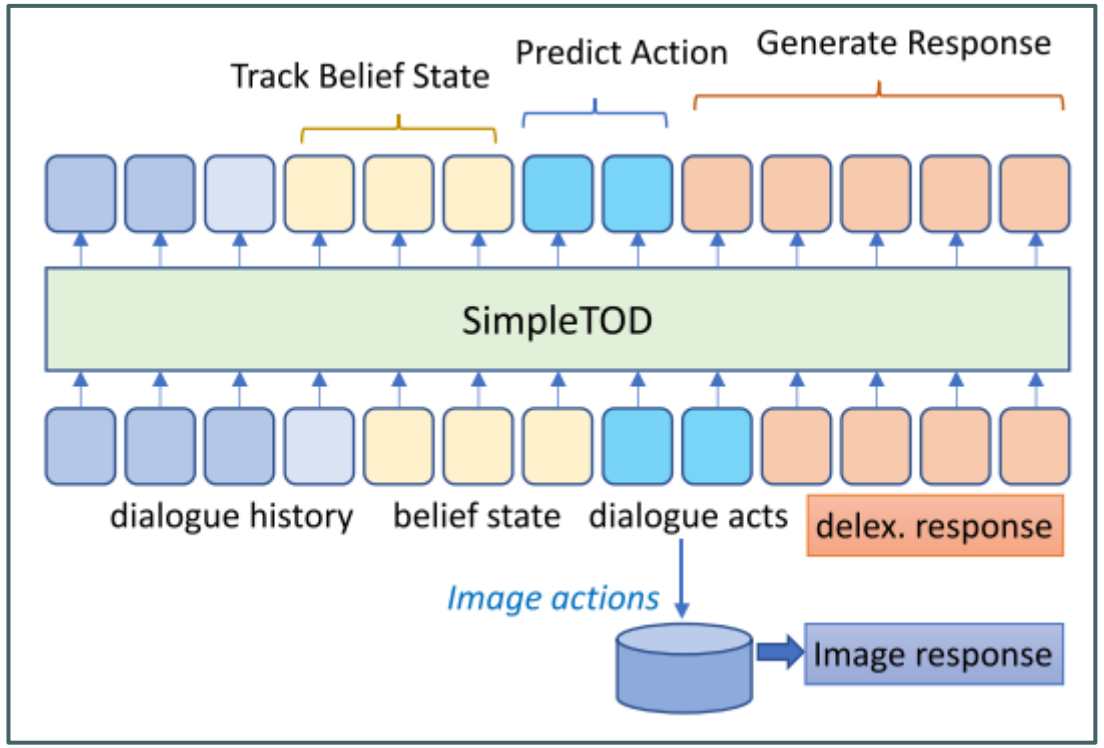
Image match

- به میزان مطابقت تصویر پیش بینی شده برای مکان صحیح اشاره دارد.

کارهای مرتبط

پژوهش‌های انجام شده مرتبط به پایان نامه

کارهای مرتبط – SimpleTOD



نمایی از مدل انطباق یافته SimpleTOD برای کامل کردن تکلیف‌های مختلف.

مدل انطباق‌یافته SimpleTOD

- انجام تکلیف‌های مختلف در یک توالی
- نتایج آن بر روی MMConv موجود است.

Liao et al. “MMConv: An Environment for Multimodal Conversational Search across Multiple Domains”. (2021)

نمونه ورودی در مدل انطباق یافته‌ی SimpleTOD

<|context|> <|system|> any particular fast food you are craving? <|user|> 4 fingers crispy chicken and bubble tea sounds like the perfect combination now. <|system|> then do you have a preferred bubble tea brand? <|user|> i love gong cha! <|system|> then this shopping center plaza singapura has all that you need. <|user|> great! any things to take note of when i am there? <|system|> there's heaps of food at the basement, so check it out when you're done. <|user|> afterwards i'm also thinking of visiting a historic site. do you know where this place in the picture is? <|endofcontext|> <|belief|> img_gt bridge inform; open span basement inform; open span bubble tea inform; open span fast food inform; open span gong cha inform; open span historic site inform; open span image inform; open span mall inform; open span place inform; open span tired inform; venuename plaza singapura confirm; venueneigh central region inform <|endofbelief|> <|action|> open span image inform; venuename cavenagh bridge recommend <|endofaction|> <|response|> <|system|> this is [venuename] in the picture. <|endofresponse|>

An example for the multitask benchmark, each dialogue turn will be represented as a sequence, which contains previous user/system turns, belief, action, and delexicalized response.
The highlighted part labels, added with EfficientNet.

کارهای مرتبط – SimpleTOD

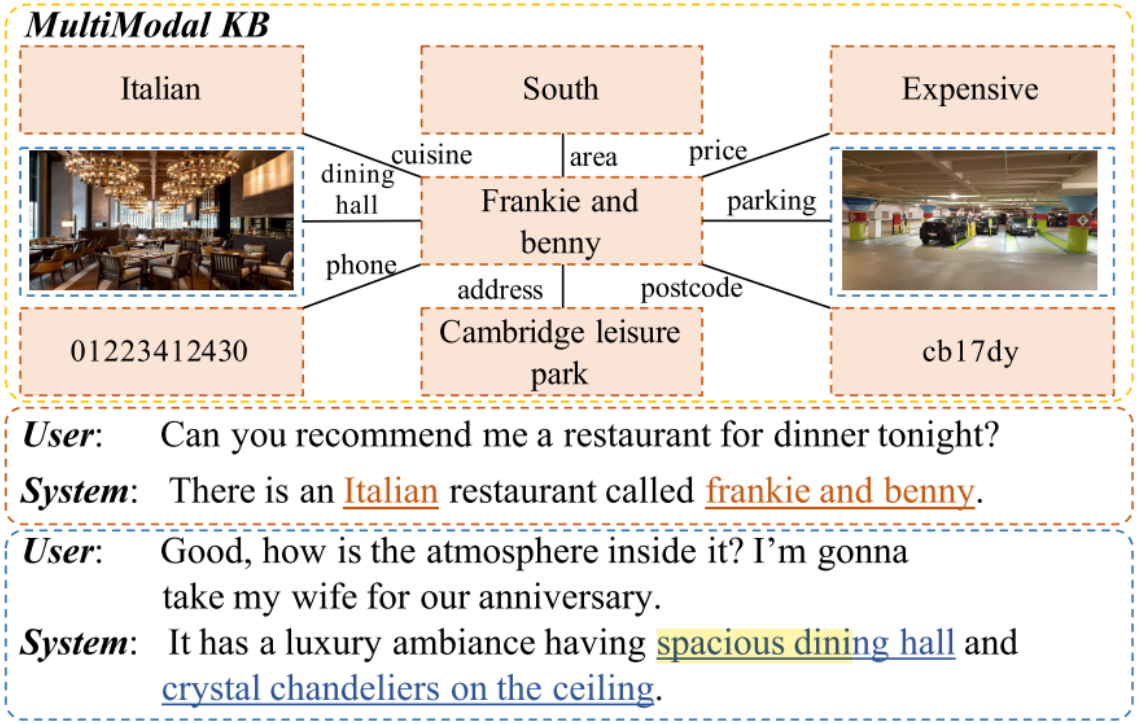
نتایج این مدل بر روی مجموعه‌داده‌های مختلف

Datasets	Joint Accuracy	Inform Rate	Success Rate	BLEU Score	Combined Score	Image Match
WOZ 2.0	0.81	77.2	68.8	18.79	91.79	–
MultiWOZ 2.0	0.57	84.4	70.1	15.01	92.26	–
MultiWOZ 2.1	0.56	85.0	70.5	15.23	92.98	–
MMConv	0.28[2]	14.6[1]	9.2[1]	20.30	32.20	0.02

[1] We evaluate on predicted agent's action results. At least one exact venue should be correct to be count as informed.
[2] Here we exclude the effect of flexible open span here .

Liao et al. "MMConv: An Environment for Multimodal Conversational Search across Multiple Domains". (2021)

کارهای مرتبط – UniMF



مجموعه داده MMDialKB:

استفاده از MultiWOZ و جستجوی نام مکان‌ها در گوگل و اضافه کردن تصویر آن به مجموعه داده

رویکرد:

استفاده از پایگاه دانش خارج از مدل و استخراج ویژگی‌های متنی (بخش نارنجی در تصویر) و ویژگی‌های تصویر (بخش آبی در تصویر)

Yang, et al. *"UniMF: A Unified Framework to Incorporate Multimodal Knowledge Bases into End-to-End Task-Oriented Dialogue Systems"*. IJCAI 2021 .

کارهای مرتبط – SIMMC

Situated and Interactive Multimodal Conversations

Seungwhan Moon*, Satwik Kottur*, Paul A. Crook†, Ankita De†, Shivani Poddar†
Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami
Eunjoon Cho, Rajen Subba, Alborz Geramifard

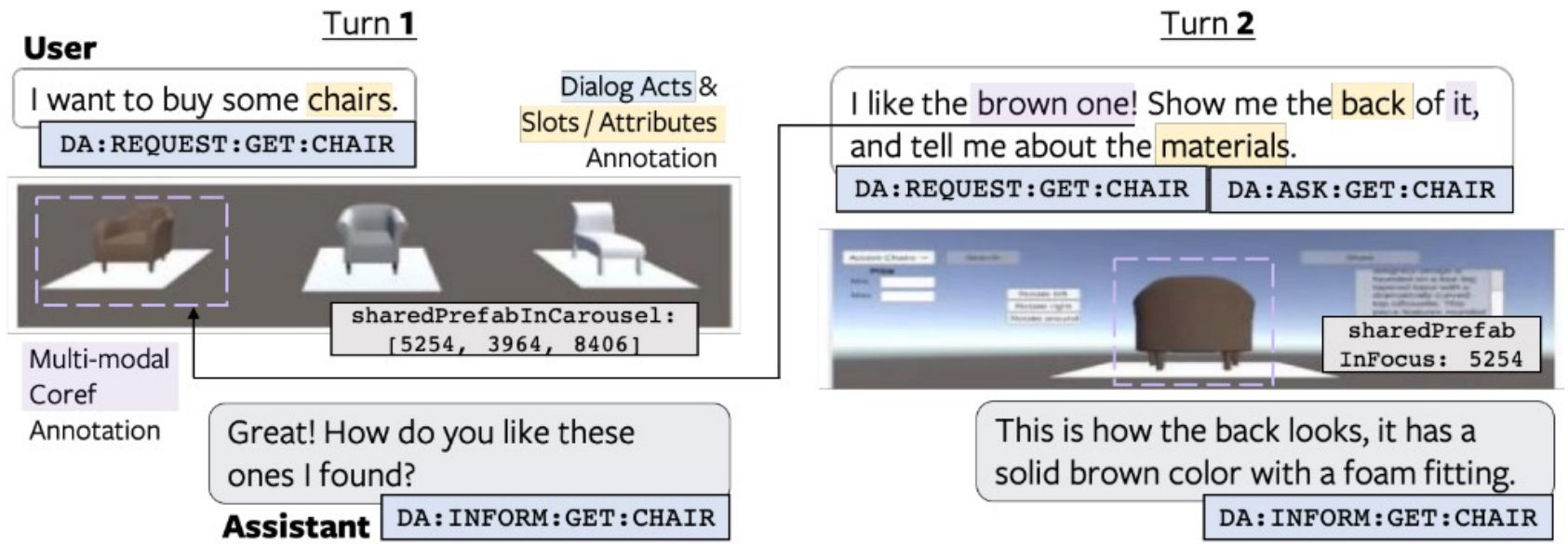
Facebook
✉ simmc@fb.com

Abstract

Next generation virtual assistants are envisioned to handle multimodal inputs (*e.g.*, vision, memories of previous interactions, and the user's utterances), and perform multimodal actions (*e.g.*, displaying a route while generating the system's utterance). We introduce Situated Interactive MultiModal Conversations (SIMMC) as a new direction aimed at training agents that take multimodal actions grounded in a *co-evolving* multimodal input context in addition to the dialog history. We provide two SIMMC datasets totalling ~13K human-human dialogs (~169K utterances) collected using a multimodal Wizard-of-Oz (WoZ) setup, on two shopping domains: (a) furniture – grounded in a shared virtual environment; and (b) fashion – grounded in an evolving set of images. Datasets include multimodal context of the items appearing in each scene, and contextual NLU, NLG and coreference annotations using a novel and unified framework of SIMMC *conversational acts* for both user and assistant utterances.

Finally, we present several tasks within SIMMC as objective evaluation protocols, such as structural API prediction, response generation, and dialog state tracking. We benchmark a collection of existing models on these SIMMC tasks as strong baselines, and demonstrate rich multimodal conversational interactions. Our data, annotations, and models are publicly available [1]

کارهای مرتبط – SIMMC



تصویر یک گفتگو بین کاربر و دستیار در simmc

Moon, S., Kottur, S., Crook, P., De, A., Poddar, S., Levin, T., … Geramifard, A. (2021). Situated and Interactive Multimodal Conversations.

قدم‌های بعدی

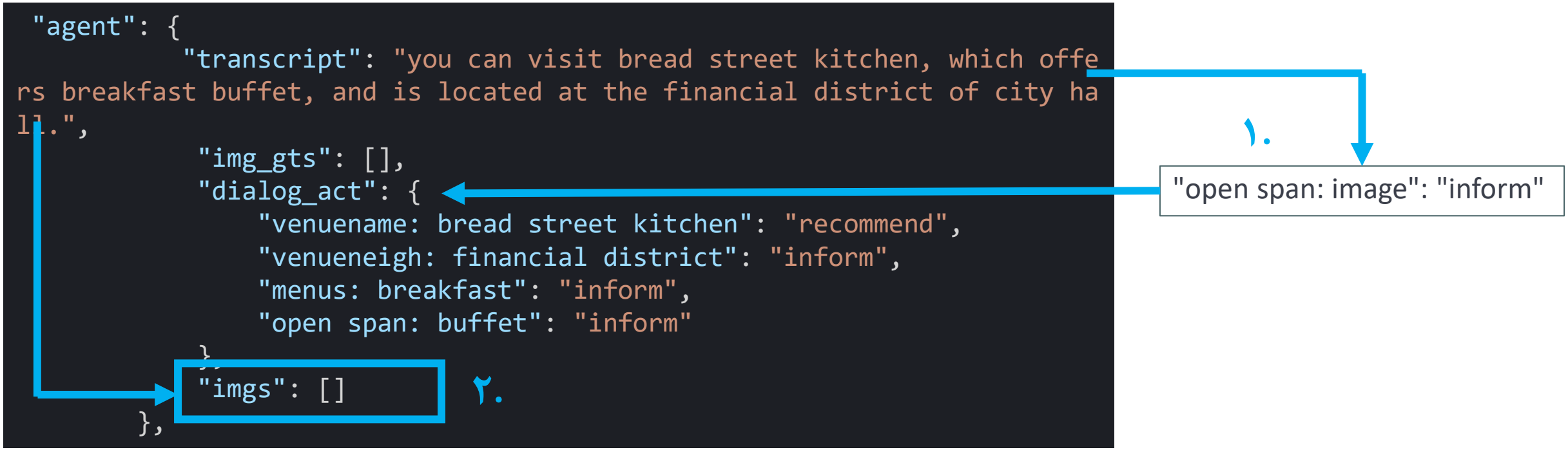
ایده و روش پیشنهادی برای حل چالش‌های مسئله

غنی سازی گفتگو

دو بخش اصلی برای غنی سازی گفتگو:

۱. تعیین زمان مناسب در بین گفتگو برای نمایش تصویر (Action prediction)

۲. پیدا کردن تصویر مرتبط به متن گفتگو از پایگاه داده یا اینترنت



غنی سازی گفتگو

تعریف بخش اول مسئله:

$$D = \{(S_i, \quad U_i, \quad M_i, A_i)\}_{i=1}^{N}$$

یک گفتگو (D) از نوبت‌های مختلف (i) تشکیل شده است.

- A_i: عمل سیستم (با API Call یا بدون آن)
- M_i: محتوای مالتی مدیا (تصویر)
- U_i: سخنان کاربر
- S_i: سخنان سیستم

غنی سازی گفتگو

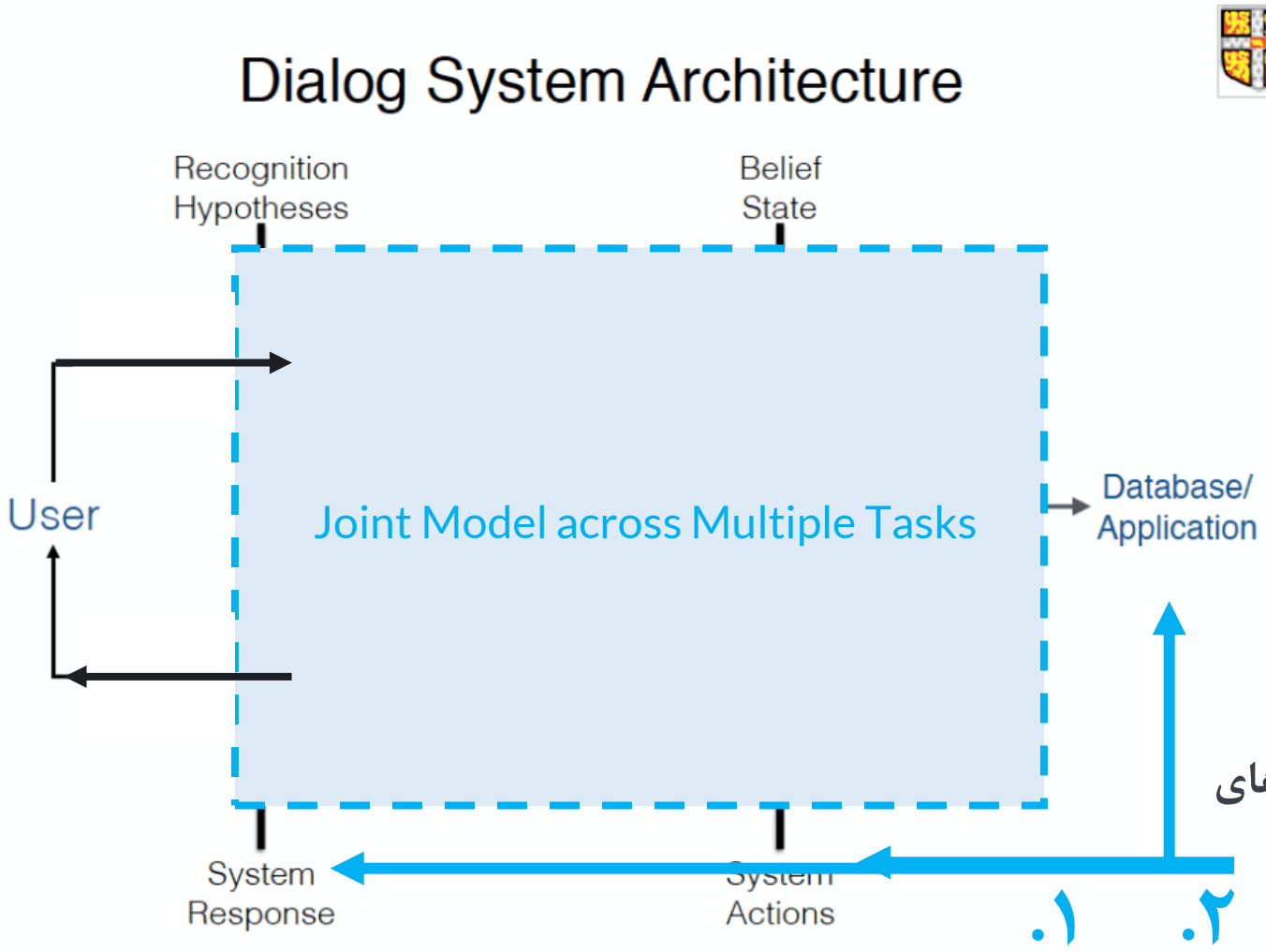
تعریف بخش اول مسئله:

$$D = \{(S_i, \quad U_i, \quad M_i, A_i)\}_{i=1}^{N}$$

فرض کنیم اگر در مرحله tام از گفتگو باشیم (داریم U_t):

- تاریخچه گفتگو تا مرحله t-1 را در اختیار داریم: $H_t = (S_i, U_i)_{i=1}^{t-1}$
- همچنین باورهای سیستم (عملیات انجام شده) تا مرحله t-1 را نیز در اختیار داریم (belief state)
- حال با توجه به سخن کاربر و محتوای مالتی مدیا در مرحله tام (M_t و U_t) باید عملیات سیستم را بدست آوریم (A_t)

غنی سازی گفتگو



معماری متداول در سیستم‌های گفتگو

سیستم‌های جدیدتر معمولا برپایه مدل‌های مشترک ساخته می‌شوند.

غنی کردن سیستم های مشترک حال حاضر

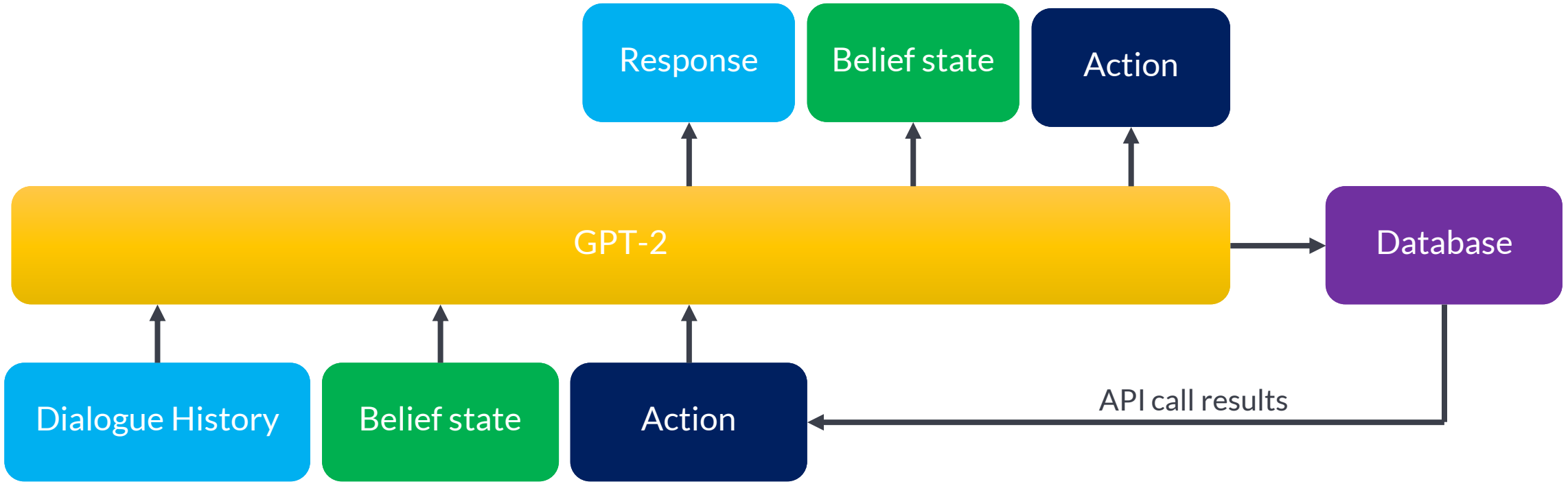
تصویر برگرفته شده از سایت https://syncedreview.com/

نحوه نمایش مدل

تاریخچه گفتگو	Belief state	Action	پاسخ به همراه مالتی‌مدیا
Ut: hi I am looking for an authentic English restaurant that serves breakfast buffet. …	"open span: authentic": "inform", "menus: breakfast": "inform"	"venuename: bread street kitchen": "recommend“, "open span: image": "inform“ **API call:** Sending a query to find an image of the recommended venue	**Delexicalized response:** St: you can go to [venuename] located in the financial district area. **Multimodal context**

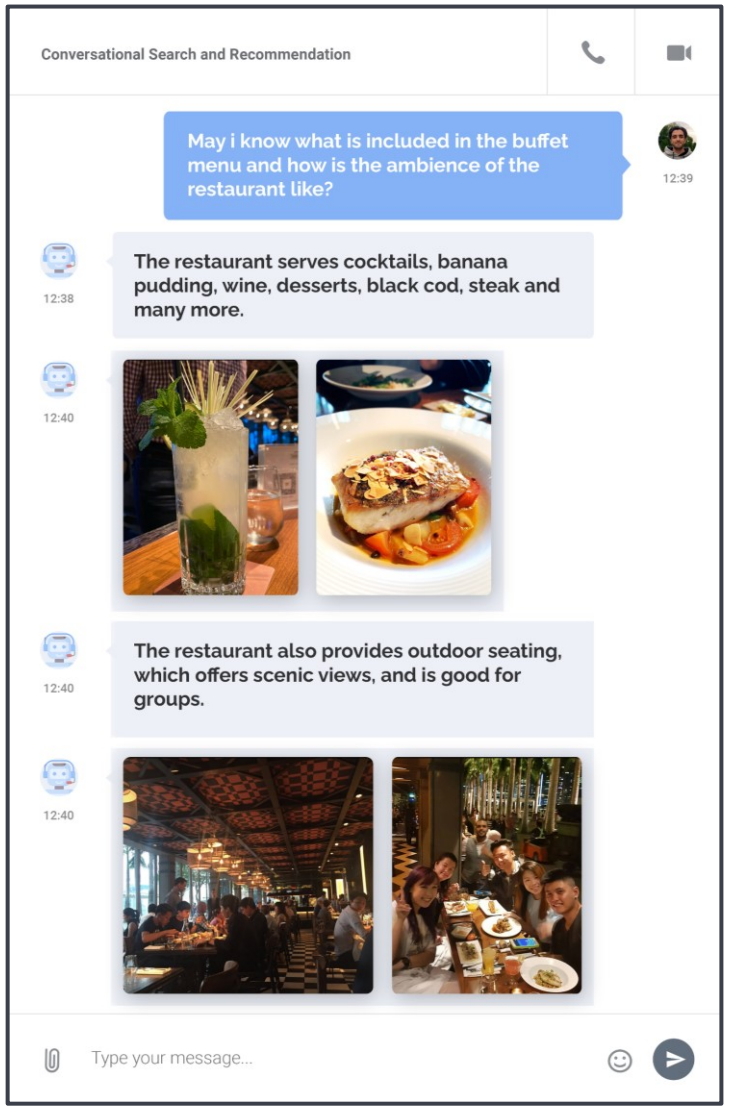
<|context|> <|user|> hi I am looking for an authentic English restaurant that serves breakfast buffet. <|system|> you can go to bread street kitchen located in the financial district area. <|endofcontext|> <|belief|> "open span: authentic": "inform","menus: breakfast": "inform<|endofbelief|> <|action|> "venuename: bread street kitchen": "recommend“, "open span: image": "inform",<|endofaction|> <|response|> <|system|> you can go to [venuename] located in the financial district area. <|endofresponse|>

چهارچوب



نمایی از چهارچوب مدل در مرحله t

جمع بندی



تصویر از نویسنده - ساخته شده بر اساس مجموعه داده MMConv

- نیاز امروز کاربران به گفتگوی مالتی مدیا و ناکافی بودن سیستم‌های متنی.
- بیشتر مجموعه داده‌های سیستم‌های گفتگو بر پایه متن هستند اما مجموعه داده‌های جدیدی به شکل مالتی مدیا در حال به وجود آمدن هستن.
- سیستم های گفتگوی متداول، بر روی مجموعه داده‌های مالتی مدیا نتایج ضعیفی دارند و نیاز به سیستم‌های جدیدتری هست.
- با افزودن تصویر به مکالمه در زمان درست، و نمایش تصاویر مرتبط می‌توان گفتگوی غنی‌تری برای کاربر ایجاد کرد.

سپاس

آیا سوال دارید؟