طراحی یک سامانه گفتگوگر وظیفه‌گرا مشترک مبتنی بر یادگیری تقویتی عمیق

محمدجواد نصری‌لوشانی[۱]، جواد سلیمی‌سرتختی[۲] و حسین ابراهیم‌پورکومله[۲]

[۱] دانشجوی کارشناسی ارشد، گروه مهندسی کامپیوتر، دانشکده مهندسی برق و کامپیوتر، دانشگاه کاشان، کاشان، ایران
mohammad.j.nasri@gmail.com

[۲] استادیار، گروه مهندسی کامپیوتر، دانشکده مهندسی برق و کامپیوتر، دانشگاه کاشان، کاشان، ایران
{salimi, ebrahimpour}@kashanu.ac.ir

چکیده‌ـ ساخت سامانه‌های گفتگوگر در سال‌های اخیر توجه زیادی به خود جلب‌کرده است. دسته‌ای از این سامانه‌ها، سامانه‌های گفتگوگر وظیفه‌گرا هستند که هدفشان رساندن انسان به مقصودش با انجام گفتگو در یک حوزه خاص می‌باشد؛ مثلاً رستوران. این سامانه‌ها از بخش‌های مختلفی تشکیل می‌شوند که اگر دو یا چند بخش، همزمان توسعه‌داده شوند، سامانه‌ی مشترک (Joint) نامیده می‌شود. یکی از روش‌هایی که برای توسعه این سامانه‌ها استفاده می‌شود، روش یادگیری تقویتی عمیق است. در یادگیری تقویتی عمیق، عامل که شبکه عصبی است با تعامل با محیط (کنش) در حالت‌های مختلف و دریافت پاداش از آن، آموزش می‌بیند. همچنین، در شروع یادگیری، عامل تعدادی کنش بصورت تصادفی انجام می‌دهد و به مرور زمان از دانشی که بدست آورده، استفاده می‌کند. در این مقاله، برای اینکه حالت مناسبی از محیط گفتگو ایجاد شود، از چسباندن نمایش جمله آخرین پیام ربات و انسان، استفاده‌شده است. همچنین، تابعی جدید برای کاهش احتمال انجام کنش تصادفی، بکارگرفته‌شده است. برای ارزیابی و مقایسه عملکرد روش ارائه‌شده با دو سامانه گفتگوگر دیگر، از شبیه‌ساز گفتگو در حوزه رستوران استفاده شده‌است. روش ارائه‌شده، بیشینه پاداش ۰/۲۹۹۳۷ را در ۲۷,۹۰۰ گام گفتگو بدست می‌آورد که نسبت به دو روش دیگر، با تعداد گفتگوهای کمتر، پاداش بیشتری بدست آورده است.

کلمات کلیدی- نمایش جمله، حوزه رستوران، مدیریت گفتگو، DQL

۱. مقدمه

یک سامانه گفتگوگر[1] یا دستیار مجازی، قادر به گفتگوی معنادار مبتنی بر متن (یا صوت) با کاربران انسانی است. یکی از اصلی ترین کاربردهای چنین سامانه‌هایی، محیط‌های خدمات رسانی به مشتری است که یک عامل مجازی جایگزین انسان می‌شود و هزینه‌های مرکز تماس و خدمات پشتیبانی را کاهش می‌دهد.

سامانه‌های گفتگوگر را می‌توان به دو دسته سامانه‌های وظیفه‌گرا و سامانه‌های غیر وظیفه‌گرا، دسته‌بندی کرد. هدف سامانه‌های وظیفه‌گرا، کمک به کاربر برای کامل کردن وظایف معینی می‌باشد؛ برای مثال تهیه بلیط هواپیما و یا بکارگیری در رستوران. از سوی دیگر، سامانه‌های غیر وظیفه‌گرا روی گفتگو با انسان در حوزه‌های باز تمرکز دارند تا واکنش‌های صحیحی برای گفتگو با انسان ایجاد کنند؛ همانند یک ربات چت.

نحوه عملکرد سامانه‌های گفتگوگر وظیفه‌گرا به این صورت می‌باشد که این سامانه‌ها ابتدا پیام ارسالی انسان را درک می‌کنند (بخش فهم زبان طبیعی) و سپس آن را به عنوان یک حالت داخلی بیان می‌کنند و با توجه به سیاست و حالت گفتگو، کُنشی اتخاذ می‌کنند (بخش مدیریت گفتگو) و در نهایت این کنش به شکل یک زبان طبیعی در آورده می‌شود (بخش تولید زبان طبیعی) و به انسان نمایش داده می‌شود.

گروهی از روش‌ها برای ساخت بخش‌های مختلف سامانه‌های گفتگوگر وظیفه‌گرا، روش‌های یادگیری ماشین هستند که به سه دسته یادگیری نظارت‌شده، یادگیری بدون نظارت و یادگیری تقویتی، تقسیم می‌شوند. سامانه‌های گفتگوگر نیاز به پیکره‌های بزرگ برای ساختن مدل‌های کارآمد دارند. از طرفی، روش‌های یادگیری تقویتی می‌توانند با استفاده از پیکره‌های کوچک شروع به یادگیری کنند و با گذشت زمان و تعامل با محیط، یادگیری انجام دهند و عملکرد خود را بهبود ببخشند. همچنین، روش‌های

[1] Dialogue System

یادگیری ماشین، نیاز به استخراج ویژگی از نمونه‌های آموزشی دارند اما روش‌های یادگیری عمیق می‌توانند خودشان با استفاده از ورودی که به آن‌ها داده می‌شود، استخراج ویژگی کنند.

رویکردهای زیادی برای ساخت سامانه‌های گفتگوگر وظیفه‌گرا ارائه‌شده است. در رویکردهای قدیمی‌تر، این سامانه‌ها در یک ساختار کلی به نحوی ساخته می‌شدند که هر بخش جداگانه طراحی، آموزش و ارزیابی می‌شد. طراحی جداگانه هر بخش، منجر به تکثیر خطا می‌شود؛ زیرا عملکرد هر بخش به خروجی بخش قبلی بستگی دارد [۱]. در سال‌های اخیر، بیشتر محققان روی رویکردهای قابل آموزش مشترک[2] کار کرده‌اند که دو یا چند بخش، ادغام‌شده و بصورت مشترک چند وظیفه را انجام می‌دهند.

در این مقاله، یک سامانه گفتگوگر وظیفه‌گرا مشترک مبتنی بر یادگیری تقویتی عمیق، طراحی‌شده است که انسان و ربات بتوانند گفتگویی از نوع متن داشته باشند و پس از اتمام گفتگو، انسان به مقصود خود برسد.

ساختار مقاله در ادامه به این صورت تدوین‌شده است: در بخش دوم مفاهیم پایه توضیح‌داده می‌شود. در بخش سوم، کارهای مرتبط در سامانه‌های گفتگوگر مشترک بررسی می‌شوند. در بخش چهارم، روش پیشنهادی ما برای ساخت این سامانه‌ها ارائه می‌شود. در بخش پنجم نیز، عملکرد روش پیشنهادی ارزیابی می‌شود و در بخش ششم، نتیجه‌گیری بیان می‌شود.

۲. مفاهیم پایه

در این بخش مفاهیم پایه مطرح می‌شوند. ابتدا طبقه‌بندی و یادگیری تقویتی، بیان می‌شوند. در ادامه، یادگیری عمیق توضیح‌داده می‌شود. سپس روش‌های نمایش کلمه و نمایش جمله معرفی می‌شوند. بخش‌های مختلف سامانه‌های گفتگوگر وظیفه‌گرا نیز در همین بخش بیان می‌شوند.

۲-۱- *طبقه‌بندی و الگوریتم بیز ساده*[3]

یک دسته از روش‌های یادگیری ماشین، روش‌های طبقه‌بندی هستند که بر اساس برچسب نمونه‌های آموزشی، یاد می‌گیرند تا با استفاده از ویژگی‌های نمونه‌ها، آنها را طبقه‌بندی کنند. روش‌های گوناگونی برای طبقه‌بندی ارائه‌شده که یکی از این روش‌ها، الگوریتم بیز ساده است که کاربرد بسیاری در پردازش زبان‌های طبیعی دارد. در این روش، از قضیه بیز و فرض استقلال بین متغیرها برای طبقه‌بندی استفاده می‌شود.

۲-۲- *یادگیری تقویتی و یادگیری Q*

برخلاف روش‌های طبقه‌بندی که مطرح شد، در یادگیری تقویتی خبری از نمونه‌های آموزشی برای یادگیری وجود ندارد و عامل (یادگیر تقویتی) با استفاده از تعامل با محیط، یعنی انجام کنش a و دریافت پاداش r، یادگیری انجام می‌دهد. حالت s، موقعیت کنونی عامل در محیط است. کنش‌های مجاز، مجموعه کنش‌هایی است که عامل می‌تواند در حالت کنونی اتخاذ کند؛ و بازخورد محیط به کنش سیستم، در قالب پاداش مطرح می‌شود. سیاست نیز یعنی به ازای هر حالت، چه کنشی اتخاذ شود.

یکی از روش‌های معروف در یادگیری تقویتی، یادگیری Q است که نگاشتی از زوجِ حالت و کنش را به مقادیری که ارزش‌ـQ نامیده می‌شوند، می‌برد. این مقادیر تحت عنوان یک جدول بنام جدول‌ـQ مطرح شود که می‌توان با استفاده از آن، سیاست مشخصی را برای انجام کنش‌های مطلوب در حالت‌های مختلف محیط پیدا کرد. مقادیر این جدول در فرآیند یادگیری با استفاده از رابطه (۱) بروز می‌شوند.

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha(R_{t+1} + \gamma \max_{a'} Q(s_{t+1},a') - Q(s_t,a_t)) \quad (۱)$$

که t بیانگر لحظه، a'کنش بدست آمده از بیشترین مقدار Q، R پاداش فوری، α ضریب یادگیری و γ ضریب تخفیف می‌باشند.

۲-۳- *سیاست ε‌ـحریصانه*[4]

در یادگیری تقویتی، عامل برای شروع تعامل با محیط، ابتدا در تعدادی گام بصورت تصادفی عمل می‌کند؛ چون از پیش دانشی ندارد تا بر اساس آن تصمیم بگیرد. در سیاست ε‌ـحریصانه، عامل با احتمال از پیش تعیین‌شده ε، کنش تصادفی انجام می‌دهد؛ به این صورت که اگر ε مقدار بیشتر از یک عدد تصادفی داشته باشد. در غیر این صورت، عامل بر اساس $Q(s,a)$ یادگرفته شده، تصمیم می‌گیرد. مقدار ε بر اساس توابعی تنظیم می‌شود که مقدار آن در ابتدا بالا و با گذر گام‌های یادگیری، کاهش می‌یابد.

۲-۴- *یادگیری عمیق*

شبکه عصبی یکی از انواع روش‌های یادگیری ماشین است که از سه لایه ورودی، پنهان و خروجی تشکیل می‌شود که وقتی تعداد لایه‌های پنهان افزایش پیدا کند، آنرا یادگیری عمیق می‌نامند که از این روش‌ها بطور گسترده در حوزه متن استفاده می‌شوند.

[2] Jointly Trainable Models
[3] Naïve Bayes
[4] Epsilon Greedy Policy

۲-۵- *یادگیری تقویتی عمیق*

این روش همانطور که از نامش پیداست، دو روش یادگیری تقویتی و یادگیری عمیق را ترکیب می‌کند. در یادگیری تقویتی عمیق، بخش عامل یادگیری تقویتی با استفاده از یک شبکه عصبی، ساخته می‌شود. در یادگیری Q، بجای جدول‌ـQ از یک شبکه عصبی برای یادگیری سیاست‌ها استفاده شود، که به این روش، یادگیری Q عمیق یا DQL گفته می‌شود. ورودی این شبکه عصبی حالت محیط و خروجی آن ارزش‌ـQ است.

۲-۶- *نمایش کلمه Glove*

روش‌های نمایش کلمه، کلمات را از یک لغت‌نامه به بردارهای عددی در یک فضای پیوسته، نگاشت می‌کنند. این روش‌ها برای کلمات مشابه، بردارهای مشابه‌ای تولید می‌کنند که از این بردارها برای نمایش کلمات استفاده می‌شود. روش‌های متفاوتی برای نمایش کلمه وجود دارد که یک از آنها روش Glove است. این روش یک روش یادگیری بدون نظارت می‌باشد که یک ماتریس از شمارش کلمات در اسناد ساخته و سپس با فاکتور گیری، ماتریسی با ابعاد کمتر می‌سازد، که هر سطر، بردار نمایش یک کلمه می‌باشد.

۲-۷- *نمایش جمله MPNet*

برای نمایش جمله، می‌توان از میانگین گرفتن بردار نمایش کلمه کلمات آن جمله استفاده کرد. تولید نمایش جمله با استفاده از این روش، نتایج مطلوبی نمی‌دهد؛ چرا که این روش‌ها برای نمایش جمله ارائه نشده‌اند. از اینرو روش‌هایی پدید آمدند که جملات را به یک بردار عددی برای نمایش آنها، در فضای پیوسته نگاشت می‌کنند و نمایشی برای جملات می‌سازند. بر طبق آزمایش روی پیکره‌های بزرگ، یکی از برترین روش‌های نمایش جمله[5] روش MPNet می‌باشد که هم از تکنیک MLM و هم از تکنیک PLM استفاده می‌کند [۲]. در MLM، وظیفه مدل، تولید جمله ورودی در خروجی است؛ به این صورت که تعدادی از کلمات ورودی به مدل، ماسک شده‌اند. در PLM، جمله ورودی در خروجی ساخته می‌شود؛ به این صورت که با استفاده از جایگشت، جای تعدادی از کلمات ورودی به مدل، عوض شده‌اند.

۲-۸- *فهم زبان طبیعی*

مدل‌های فهم زبان طبیعی، اطلاعات گفتگو را برای بخش مدیریت گفتگو فراهم می‌کنند. این اطلاعات شامل شناسایی دامنه، شناسایی مقصود و شناسایی اسلات‌ها یا تشخیص مفهوم می‌باشد. برای حل مسائل تشخیص دامنه و تشخیص مقصود، می‌توان از طبقه‌بندها استفاده کرد.

۲-۹- *مدیریت گفتگو*

بیشتر مدل‌های مدیریت گفتگو تلاش می‌کنند تا گفتگو را با حالت‌های مشاهده‌شده در مدل، مدیریت کنند. این روش‌ها به دو بخش ردیابی حالت و تولید کنش یا سیاست، تقسیم می‌شوند. ردیابی حالت، حالت گفتگوی کنونی را با توجه به تاریخچه گفتگو، بیان می‌کند. بخش تولید کنش، ارتباطی روی حالت‌ها و کنش‌های سیستم می‌سازد که یادگیری تقویتی برای این امر بطور گسترده استفاده می‌شود.

۲-۱۰- *تولید زبان طبیعی*

مدل‌های تولید زبان طبیعی، کنش انتخابی در بخش مدیریت گفتگو را تبدیل به یک زبان طبیعی می‌کنند. خروجی این مدل‌ها در قالب متن یا صوت، به انسان نمایش‌داده می‌شود.

۳. مرور کارهای مرتبط

در زمینه سامانه‌های گفتگوگر وظیفه‌گرا مشترک، روش‌های زیادی ارائه‌شده است. در بخش فهم زبان طبیعی، با استفاده از شبکه‌های عصبی بازگشتی، شناسایی مقصود و تشخیص مفهوم بصورت مشترک [۳] و هر سه وظیفه فهم زبان طبیعی بصورت مشترک [۴] را انجام می‌دهند. در بخش مدیریت گفتگو، با استفاده از یادگیری تقویتی، روشی ارائه شده‌است که هر دو وظیفه مدیریت گفتگو را بصورت مشترک انجام می‌دهد [۵]. در این روش، یک سامانه پیشنهاددهنده ساخته‌شده است که با استفاده از گفتگو، یک محصول به کاربر پیشنهاد می‌دهد.

در بخش‌های فهم زبان طبیعی و مدیریت گفتگو بصورت مشترک، روش SimpleDS طراحی‌شده است که برداری جهت نمایش پیام انسان و پیام ربات با استفاده از نمایش کلمه Glove می‌سازد و سپس این بردار را به عنوان ورودی به یک شبکه Q عمیق می‌دهد تا یادگیری انجام شود [۶]. در این روش از کاهش ε با تابع خطی در سیاست εـ حریصانه استفاده‌شده و همچنین از شبیه‌ساز کاربر بجای انسان استفاده‌شده و برای تعیین کنش‌های مجاز یادگیر تقویتی (یا ربات) از بیز ساده استفاده کرده‌اند. از طرفی، این روش قابلیت گفتگوی صوتی نیز داشت و در [۷]، این سامانه برای استفاده در چند دامنه بصورت همزمان توسعه‌داده شده‌است. روش SCGSimpleDS توسعه‌ای بر روش SimpleDS می‌باشد که با استفاده از شبکه‌های خود رمزگذار، بردار نمایش پیام انسان و پیام ربات را فشرده‌کرده و با استفاده از تشخیص احساسات گفتگوی انسان و خوشه‌بندی حالات محیط با استفاده از روش K-Means، بردار ورودی به شبکه تقویتی عمیق را، بازسازی کرده‌اند [۸]. دلیل تغییر حالت محیط در این کار، تولید حالت‌های بهتر از

[5] www.sbert.net/docs/pretrained_models.html

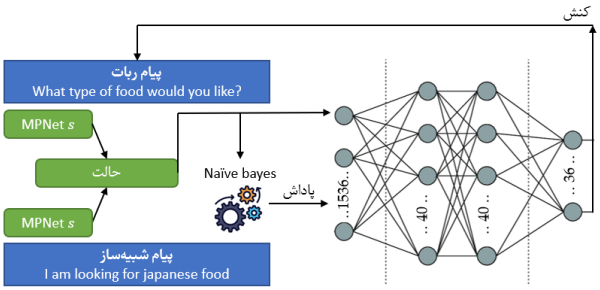
محیط می‌باشد. همچنین در این روش، با استفاده از شبکه‌های مولد تخاصمی[6]، برای آموزش بیز ساده داده‌های بیشتری تولید کرده‌اند. در روشی دیگر که مبتنی بر یادگیری تقویتی عمیق برای فهم زبان طبیعی و مدیریت گفتگو می‌باشد [۹]؛ شبکه یادگیری تقویتی ابتدا بروی یک حوزه آموزش‌دیده و سپس روی حوزه دیگری با وزن‌های مدل قبلی آموزش می‌بینند. این روش از ابتدا، مجموعه کنش‌های دو حوزه را برای هر دو مدل در نظر می‌گیرد.

در بخش تولید زبان طبیعی بصورت مشترک، با استفاده از شبکه‌های عصبی بازگشتی، مدلی طراحی‌شده است که می‌تواند هر دو وظیفه تولید زبان را انجام دهد [۱۰]. در این روش، هم از یک روش رتبه‌دهنده و هم از یک روش تولید دنباله برای تولید جملات استفاده‌شده است.

برخی روش‌ها نیز ارائه شده‌اند که تمامی بخش‌های یک سامانه گفتگوگر را در یک مدل، به انجام رسانده باشند. در [۱] با استفاده از چند شبکه عصبی بازگشتی سلسه مراتبی، روشی برای این کار ارائه شده‌است که در آن برای مدیریت گفتگو از یادگیری تقویتی استفاده‌شده است.

۴. روش پیشنهادی

روشی که برای ساخت سامانه گفتگوگر وظیفه‌گرا مشترک ارائه‌شده، مبتنی بر یادگیری تقویتی عمیق می‌باشد که برای ساخت حالت از نمایش جمله MPNet در آن استفاده‌شده است که معماری آن در شکل ۱ قابل مشاهده است. برای آموزش یادگیری تقویتی عمیق، از شبیه‌ساز کاربر استفاده‌شده که نقش انسان را برای تولید گفتگو، ایفا می‌کند. در ادامه، بخش‌های مختلف سامانه بیان خواهد شد.



شکل ۱: معماری سامانه گفتگوگر وظیفه‌گرا مشترک ارائه‌شده

۴-۱- *شبیه‌ساز کاربر*

در این سامانه گفتگوگر، بجای انسان از شبیه‌ساز کاربر استفاده می‌شود. شبیه‌ساز، جملاتش را بصورت تصادفی با تعدادی اسلات از پیش تعیین‌شده، پر می‌کند و تولید گفتگو می‌کند. برای آموزش یادگیر تقویتی، نیاز به گام‌های زیاد و نمونه‌های آموزشی می‌باشد که استفاده از شبیه‌ساز، می‌تواند این مسئله را بخوبی پوشش دهد.

۴-۲- *حالت*

برای ساخت حالت، از برداری استفاده می‌شود که از چسباندن آخرین بردار نمایش پیام ربات و پیام شبیه‌ساز که هر کدام از طریق نمایش جمله MPNet بدست آمده‌اند، ساخته می‌شود.

۴-۳- *پرونده نمایش*

از گفتگوهایی که در نمونه‌های آموزشی وجود دارد، یک پرونده نمایش ساخته خواهد شد که از آن برای آموزش طبقه‌بند بیز ساده استفاده می‌شود. در ساخت پرونده نمایش، از روی گفتگوها، بردار حالت آنها ساخته می‌شود و سپس کنش انتخابی ربات به آخر آن چسبانده می‌شود؛ از کنش ربات برای برچسب طبقه‌بند استفاده می‌شود.

۴-۴- *پاداش*

در این سامانه گفتگوگر، تابع پاداش از رابطه (۲) بدست می‌آید که از سه بخش تشکیل‌شده است. بخش Cr، از اسلات‌های تاییدشده توسط کنش کنونی، تقسیم بر تعداد کل اسلات‌هایی که باید پر شود، بدست می‌آید. بخش Dr، مقدار شباهت حالت کنونی به حالت‌های موجود در پرونده نمایش است؛ که توسط طبقه‌بند بیز ساده بدست می‌آید. بخش DL نیز مشوّق گفتگوهای کارآمدتر است.

$$R = (Cr \times \alpha) + (Dr \times (1 - \alpha)) - DL \qquad (۲)$$

۴-۵- *مقدار ε*

در این سامانه گفتگوگر، دو روش پیاده‌سازی شده‌است که روش اول از تابع خطی برای کاهش مقدار ε استفاده می‌کند که SenSimpleDS نامیده‌شده و روش دوم از تابعی دیگر که در رابطه (۳) آمده است، برای کاهش مقدار ε استفاده می‌کند که SenSimpleDS+ نامیده شده‌است.

$$\varepsilon_{new} = \max(\varepsilon_{min}, (\varepsilon \times \beta)) \qquad (۳)$$

که ε_{new} مقدار جدید و ε_{min} کمترین مقدار برای ε؛ β ضریب کاهنده و ε مقدار فعلی ε هستند.

۵. آزمایش‌ها

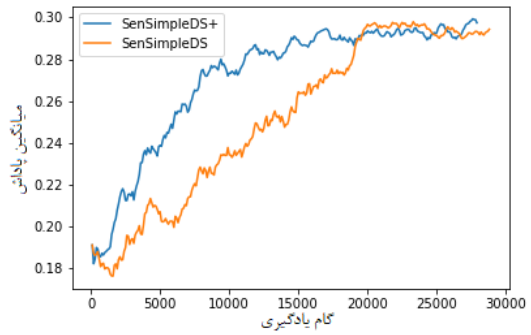
در این بخش، عملکرد سامانه گفتگوگر ارائه‌شده، ارزیابی و با دو روش SimpleDS و SCGSimpleDS مقایسه می‌شود. برای این

[6] Generative Adversarial Network (GAN)

امر، سامانه در حوزه رستوران توسعه‌یافته و از پیکره ارائه‌شده توسط روش SimpleDS که قابل دسترس است[7]، استفاده می‌شود. در این بخش همچنین، عملکرد دو روش SenSimpleDS و SenSimpleDS+ در یادگیری نیز، مقایسه می‌شوند.

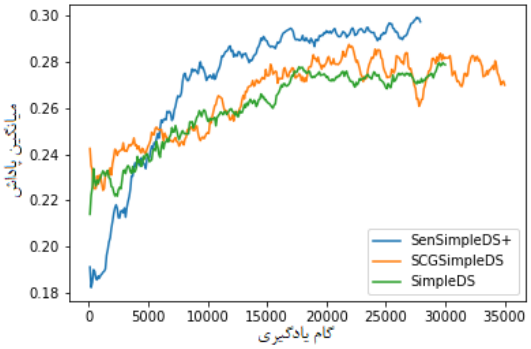
این سامانه با زبان پایتون و در محیط گوگل کولب، پیاده‌سازی شده است؛ بطوری که عامل و محیط کاملاً از هم مستقل هستند و ارتباط بین آن‌ها متن پیام، حالت، پاداش و کنش می‌باشد. شبیه‌ساز کاربر دارای ۳۳ نوع گفتگو است که چهار اسلات را با مقادیر تصادفی پر می‌کند. بردار حالت با طول ۱۵۳۶، از چسباندن بردار نمایش MPNet ربات و شبیه‌ساز بدست می‌آید. تعداد شش فایل گفتگوی کامل در پیکره وجود دارد که پرونده نمایش از روی آنها ساخته می‌شود. در تابع پاداش، در بخش Cr تعداد کل اسلات‌هایی که باید پر شود برابر با سه؛ مقدار DL برابر با ۰/۱ و مقدار α برابر با ۰/۵، در نظر گرفته شده‌است. با این مقدار α، سامانه بطور مساوی به اسلات‌های پرشده و شبیه‌بودن حالت فعلی به پرونده نمایش، اهمیت می‌دهد. برای روش SenSimpleDS+، مقدار β برابر با ۰/۹۹۹۸ در نظرگرفته‌شده است، تا مقدار ε تا جایی که می‌شود به آهستگی کاسته شود و در حدود گام یادگیری ۳۵،۰۰۰ به ε_{min} برسد. برای شبکه یادگیری تقویتی عمیق، لایه ورودی تعداد ۱۵۳۶ گره و دو لایه پنهان با ۴۰ گره، در نظر گرفته‌شده است. لایه خروجی نیز ۳۶ گره دارد که برابر با تمام کنش‌های ربات می‌باشد. همچنین، با استفاده از طبقه‌بند بیز ساده، کنش‌هایی که احتمال حداقل ۰/۰۰۱ داشته‌باشند به عنوان کنش‌های مجاز تعیین می‌شوند.

برای مقایسه دو روش SenSimpleDS و SenSimpleDS+، هر دو روش برای ۳۰۰۰ دور گفتگو با شبیه‌ساز، آموزش دیده‌اند. پاداش بدست‌آمده توسط آنها در ۱۰۰ گام یادگیری (یک گام گفتگو)، میانگین متحرک گرفته‌شده که در شکل ۲ قابل مشاهده است. از ابتدای آموزش تا قبل از گام ۲۰،۰۰۰، روش دوم پاداش بیشتری جمع‌آوری‌کرده است که دلیل آن استفاده از تابع دوم برای کاهش مقدار ε است که منجر به یادگیری سریعتر یادگیری تقویتی عمیق شده و از گام ۲۰،۰۰۰ به بعد، روش اول به میانگین پاداش بیشتری می‌رسد. روش اول و دوم، بترتیب تعداد ۲۸،۸۰۰ و ۲۷،۹۰۰ گام یادگیری را طی‌کرده‌اند که نشان می‌دهد روش دوم در گام‌های کمتری گفتگوها را به پایان رسانده که در شکل ۲ نیز مشخص است؛ دلیل آن نیز استفاده از تابع دوم، برای کاهش مقدار ε می‌باشد.



شکل ۲: پاداش جمع‌آوری شده در ۳۰۰۰ دور گفتگو توسط روش‌های ارائه شده؛ خط نارنجی SenSimpleDS و خط آبی SenSimpleDS+.

برای مقایسه عملکرد روش SenSimpleDS+ ارائه‌شده با SimpleDS و SCGSimpleDS، همه روش‌ها برای ۳۰۰۰ دور گفتگو با شبیه‌ساز، آموزش دیده‌اند و نمودار میانگین پاداش آنها برای ۱۰۰ گام یادگیری، رسم‌شده است که در شکل ۳ آورده‌شده است. دو روش SimpleDS و SCGSimpleDS، بترتیب تعداد ۳۰،۰۰۰ و ۳۲،۵۰۰ گام یادگیری را طی‌کرده‌اند که از روش SenSimpleDS+، بیشتر است. این یعنی روش SenSimpleDS+ در تعداد گفتگوهای کمتری انسان را به مقصودش می‌رساند. همچنین روش SenSimpleDS+ نسبت به دو روش دیگر، میانگین پاداش بیشتری را جمع‌آوری‌کرده است که در شکل ۳ نیز قابل مشاهده است. قبل از گام ۵۰۰۰، این روش میانگین پاداش کمتری دارد که یعنی این روش نسبت به دو روش دیگر دیرتر یادگیری انجام‌داده است، که بدلیل بزرگتر بودن اندازه بردار حالت آن می‌باشد.



شکل ۳: نتایج بدست آمده توسط روش SenSimpleDS+ (خط آبی) و مقایسه آن با سایر روش‌ها؛ خط سبز SimpleDS و خط نارنجی SCGSimpleDS.

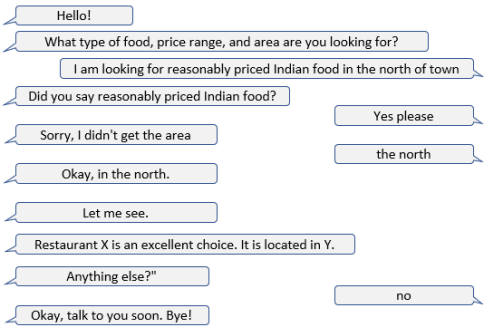
[7] github.com/cuayahuitl/SimpleDS

بیشینه پاداش جمع‌آوری‌شده روش‌های ارائه‌شده و دو روش SimpleDS و SCGSimpleDS در جدول ۱ آورده شده‌است که بیشترین مقدار آن مربوط به روش SenSimpleDS+، می‌باشد.

جدول ۱: بیشینه پاداش بدست‌آمده توسط روش‌های مختلف در ۳۰۰۰ دور گفتگو با شبیه‌ساز.

نام روش	بیشینه پاداش
SimpleDS [6]	۰/۲۷۹۷
SCGSimpleDS [8]	۰/۲۸۷۶
SenSimpleDS	۰/۲۹۷۹۵
SenSimpleDS+	**۰/۲۹۹۳۷**

در ادامه در شکل ۴، یک نمونه از گفتگوی ربات و شبیه‌ساز کاربر، توسط روش SenSimpleDS+، آورده‌شده است.



شکل ۴: نمونه از گفتگوی روش ارائه شده در زبان انگلیسی؛ سمت چپ ربات و سمت راست شبیه‌ساز.

۶. نتیجه‌گیری

در سال‌های اخیر، توسعه سامانه‌های گفتگوگر وظیفه‌گرا مشترک، بسیار مورد توجه قرار گرفته است. یکی از روش‌های توسعه این سامانه‌ها، استفاده از یادگیری تقویتی عمیق می‌باشد. در این مقاله، با استفاده از یادگیری تقویتی عمیق و شبکه MPNet که برای نمایش جمله بکار می‌رود، روشی بنام SenSimpleDS برای توسعه سامانه گفتگوگر وظیفه‌گرا مشترک ارائه شد. در این روش، برای تعیین کنش‌های مجاز یادگیری تقویتی عمیق، از الگوریتم بیز ساده استفاده شد. با اضافه کردن تابع نمایی برای کاهش ε، SenSimpleDS+ روش دیگری بود که ارائه شد. عملکرد روش‌های ارائه‌شده در حوزه رستوران، ارزیابی شد و نتایج آن با روش‌های دیگر با استفاده از گفتگو با شبیه‌ساز مقایسه شد، که نسبت به آنها، بهبود ایجاد کرده‌اند. استفاده از نمایش جمله برای ساخت بردار حالت، توانست نمایش بهتری از محیط بسازد و در نتیجه عامل یادگیری تقویتی عمیق، پاداش بیشتری دریافت کرد؛ که نشان از یادگیری بهتر آن دارد. استفاده از تابع جدید برای کاهش ε نیز، بخوبی تناسبی میان استفاده از دانش کنونی و بدست آوردن دانش جدید (با کنش تصادفی)، ایجاد کرد، که منجر به یادگیری سریعتر عامل یادگیر تقویتی شد.

مراجع

[۱] B. Liu, G. Tur, D. Hakkani-Tur, P. Shah, and L. Heck, "Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems," *arXiv preprint arXiv:1804.06512*, 2018.

[۲] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu" ,Mpnet: Masked and permuted pre-training for language understanding," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16857-16867, 2020.

[۳] X. Zhang and H. Wang, "A joint model of intent determination and slot filling for spoken language understanding," in *IJCAI*, 2016, vol. 16, no. 2016, pp. 2993-2999 .

[۴] Y. Shi, K. Yao, H. Chen, Y.-C. Pan, M.-Y. Hwang, and B. Peng, "Contextual spoken language understanding using recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015: IEEE, pp. 5271-5275 .

[۵] Z. Yan, N. Duan, P. Chen, M. Zhou, J. Zhou, and Z. Li, "Building task-oriented dialogue systems for online shopping," in *Thirty-first AAAI conference on artificial intelligence* ,2017 .

[۶] H. Cuayáhuitl, "Simpleds: A simple deep reinforcement learning dialogue system," in *Dialogues with social robots*: Springer, 2017, pp. 109-118.

[۷] H. Cuayáhuitl, S. Yu, A. Williamson, and J. Carse, "Scaling up deep reinforcement learning for multi-domain dialogue systems," in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017: IEEE, pp. 3339-3346 .

[۸] D. Zahra and S. S. Javad, "Design And Development Of A Dialogue System Using Deep Reinforcement Learning," in *12th International Conference on Information Technology, Computer and Telecommunication*, 2021 .

[۹] V. Ilievski, C. Musat, A. Hossmann, and M. Baeriswyl, "Goal-oriented chatbot dialog management bootstrapping with transfer learning," *arXiv preprint arXiv:1802.00500*, 2018.

[۱۰] O. Dušek and F. Jurčíček, "Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings," *arXiv preprint arXiv:1606.05491*, 2016.