یادگیری تقویتی عمیق برای تولید دیالوگ

چکیده

مدلهای عصبی اخیر تولید دیالوگ برای ایجاد پاسخ ها برای عامل های مکالمه بسیار نویدبخش بوده است، ولیکن میل به نزدیک بینی دارد بطوریکه بیان را یک بار پیشگویی می کند، درحالیکه تاثیر آنها را بر نتایج آتی نادیده می انگارد. مدلسازی جهت آتی دیالوگ در ایجاد دیالوگ منسجم و جالب امری حیاتی و مهم است، نیازی که مدلهای NLP قدیمی دیالوگ را منجر به استنباط طبق یادگیری پسخورد نموده است. در این مقاله، ما نشان داده ایم که چگونه این اهداف با هم ترکیب شده و یادگیری پسخورد عمیق را برای مدلسازی پاداش آتی در دیالوگ محاوره ای بکارمی بندد. این مدل دیالوگ ها را بین دو عامل مجازی با استفاده از روشهای گرادیانی سیاستگزاری شبیه سازی می کند تا نتایجی را پاداش دهد که سه خصوصیت مکالمه ای مفید را نمایش می دهند: اطلاع رسانی، انسجام و سهولت پاسخ دهی (که به عملکرد آینده نگری مربوط می شود). ما مدل خودمان را در زمینه تنوع، طول مدت و قضاوت های انسانی ارزیابی کرده ایم که نشان می دهد الگوریتم مطرح شده باعث ایجاد پاسخ های تعاملی تر و مدیریت شکوفایی مکالمه پایدارتر در شبیه سازی دیالوگ می شود. این کار اولین مرحله به سمت یادگیری یک مدل مکالمه عصبی را مبتنی بر موفقیت طولانی مدت دیالوگ نشان می دهد.

1-مقدمه

تولید پاسخ عصبی همچنان مورد علاقه بوده است. مدل توالی به توالی LSTM (SEQ2SEQ) یک نوع مدل تولید عصبی است که احتمال تولید یک پاسخ را با درنظرگیری نوبت دیالوگ قبلی به حداکثر می رساند. این شیوه باعث اضافه کردن زمینه غنی می شود که بین نوبت های دیالوگ متوالی متناظر است به شیوه ای که برای مثال با مدلهای دیالوگ مبتنی بر MT امکانپذیر نیست.

علی رغم موفقیت مدلهای SEQ2SEQ در تولید دیالوگ، دو مسئله پدیدار شده سات: اول اینکه، مدلهای SEQ2SEQ با پیشگویی نوبت دیالوگ بعدی در یک زمینه مکالمه معین با استفاده از تابع عینی تخمین احتمال ماکزیمم MLE آموزش می بینند. ولیکن، روشن نیست که به چه خوبی MLE هدف دنیای واقعی تدوین محاوره را تخمین می زند: تعلیم یک ماشین برای مکالمه با انسانها ضمن ارائه فیدبک جالب و متنوع و آموزنده که کاربران

را مشغول نگه می دارد. یک مثال منسجم این است که مدلهای SEQ2SEQ تمایل دارند که پاسخ های عمومی تری را نظیر «نمی دانم» صرفنظر از ورودی ایجاد کنند. این امر می تواند به فراوانی بالای پاسخ های عمومی یافت شده در مجموعه آموزشی و قابلیت سازگاری آنها با مجموعه متنوع زمینه های مکالمه ای نسبت داده شود. با اینحال، «نمی دانم» ظاهرا یک اقدام خوبی برای اتخاذ می باشد، چرا که مکالمه را خاتمه می دهد.

مسئله متداول دیگری که در دو مکالمه نمونه در سمت چپ جدول 1 نمایش داده شده است، این است که سیستم در یک حلقه بینهایت پاسخ های تکراری گیر کرده است. این امر به دلیل عدم توانایی مدلهای SEQ2SEQ مبتنی بر MLE برای توجیه تکرار می باشد. در مثال 2 (پایین سمت چپ)، مکالمه به یک حلقه بی نهایت بعد از سه نوبت می افتد، که هر دو عامل ایجاد کننده بیانهای کسل کننده عمومی نظیر «نمی دانم درباره چه صحبت می کنید و نمی دانم شما چه می گویید.» می باشند. با نگاه به کل مکالمه، بیان تبدیل به اتخاذ یک عمل بد می شود چرا که هیچ شیوه ای را برای ادامه مکالمه ارائه نمی دهد.[1]

مدلهای یادگیری پسخورد مطرح شده	مدل اطلاعات دوجانبه پایه
الف-داری کجا می ری؟ (1) ب-دارم به ایستگاه پلیس می رم. (2) الف-منم باهات می یام. (3) ب-نه، نه، نه، نه، تو هیچ جا نمی یای. (4) الف-چرا؟ (5) ب-می خوام اینجا بمونی. (6) الف-نمی دونم درباره چی حرف می زنی؟ (7) ...	الف-داری کجا می ری؟ (1) ب-دارم می رم دستشویی. (2) الف-بعدا می بینمت.(3) ب-بعدا می بینمت.(4) الف-بعدا می بینمت.(5) ب-بعدا می بینمت.(6) ... ...
الف-چند سالته؟ (1) ب-شانزده سالمه. چرا می پرسی؟ (2) الف-فکر کردم 12 سالته. (3) ب-چی باعث شد اینجوری فکر کنی؟ (4) الف-نمی دونم درباره چی صحبت می کنی. (5) ب-نمی دونی درباره داری چی می گی. (6)	الف-چند سالته؟ (1) ب-شانزده سالمه. (2) الف-16 سال؟ (3) ب-نمی دونم درباره چی صحبت می کنی. (4) الف-نمی دونی درباره داری چی می گی. (5) ب-نمی دونم درباره چی صحبت می کنی. (6) الف-نمی دونی درباره داری چی می گی. (7) ...

[1] قانون مشابهی اغلب در کمدی بداهه گویی ارائه شده است:
https://en.wikipedia.org/wiki/Yes,_and...

جدول 1-ستون سمت چپ: شبیه سازی دیالوگ بین دو عامل با استفاده از یک کدگذاری کننده-کدگشاینده LSTM چهار لایه که روی پایگاه داده های OpenSubtitles امتحان شده است. اولین دور (فهرست 1) توسط نویسندگان به شکل ورودی وارد گردیده است. بعد این دو عامل به نوبت مکالمه کرده و نوبت مکالمه ایجاد شده قبلی عامل دیگر را به شکل ورودی درنظر گرفتند. خروجی (نتیجه) با استفاده از مدل اطلاع رسانی متقابل ایجاد شده است که در آن فهرست n بهترین ابتدا با استفاده از جستجوی پرتویی مبتنی بر $p(t|s)$ بدست می آید و با ترکیب خطی احتمالات عقب گرد $p(s|t)$ رتبه بندی مجدد می شود که در آن t و s به ترتیب نمایانگر هدف و منابع می باشند. **ستون سمت راست:** مکالمه با استفاده از مدل یادگیری پسخورد مطرح شده شبیه سازی شده است. مدل جدید آینده نگرانه بیشتری دارد (سوالاتی نظیر چرا می پرسی؟ و پیشنهاداتی نظیر منم باهات می یام.) و به مدت طولانی تری باقی می ماند قبل از اینکه در چاله های سیاه مکالماتی گیر بیافتد.

این چالش ها حاکی از آنست که ما نیاز به یک چارچوب مکالماتی داریم که توانایی 1)ترکیب پاداش های تعریف شده ابداع کننده و تقلید بهتر هدف حقیقی تولید محاوره و 2) مدلسازی تاثیر طولانی مدت یک پاسخ ایجاد شده در یک مکالمه مداوم را دارد.

برای دستیابی به این اهداف، ما طبق بینش های یادگیری پسخورد استنباط انجام داده ایم که به طور گسترده ای در سیستم های محاوره ای MDP و POMDP بکار بسته شده است (بخش مطالعه مرتبط را برای جزئیات ببینید). ما یک روش ایجاد یادگیری پسخورد عصبی RL را معرفی کرده ایم که می تواند پاداش های طولانی مدت را بهینه سازی کند که توسط ابداع گران سیستم طراحی شده است. مدل ما از معماری کدگذاری کننده-کدگشایی کننده به شکل ستون فقرات آنها استفاده می کند، و مکالمه را بین دو عامل مجازی برای بررسی فضای اقدامات احتمالی حین یادگیری به حداکثررسانی پاداش مورد انتظار شبیه سازی می کند. ما تخمین های سلسله مراتبی ساده ای را برای پاداش دهی تعریف کرده ایم که مکالمات خوب را مشخصه سازی می کند: مکالمات خوب آینده نگرانه یا تعاملی (یک نوبت مکالمه به نوبت بعدی پیشنهاد می کند)، آموزنده و منسجم هستند. پارامترهای یک RNN کدگذاری کننده-کدگشاینده یک سیاسگزاری را روی یک فضای عملکرد بی نهایت شامل کلیه بیانهای محتمل تعریف می کند. این عامل یک سیاستگزاری را با بهینه سازی پاداش بلندمدت معین شده توسط سازنده از روی شبیه سازی های مکالمات در حال انجام با استفاده از روشهای گرادیان سیاستگزاری به جای هدف MLE تعریف

شده در مدلهای استاندارد SEQ2SEQ یاد می گیرد. بااینحساب، مدل ما قدرت سیستم های SEQ2SEQ را به کار می گیرد تا معانی معنایی انشایی بیان ها را با قدرت یادگیری پسخورد در بهینه سازی برای اهداف بلند مدت طی یک مکالمه یاد بگیرد. نتایج آزمایشی (نتایج نمونه گیری شده در پانل سمت راست جدول 1) نشان می دهد که روش ما یک دیالوگ پایدارتری را شکوفا می سازد و مدیریت می کند تا پاسخ های تعاملی تری را نسبت به مدلهای استاندارد SEQ2SEQ که با استفاده از هدف MLE آموزش دیده است، ایجاد کند.

2-تحقیقات مرتبط

تلاشها برای ایجاد سیستم های دیالوگ آماری به دو دسته تقسیم بندی می شود.

اولی تولید دیالوگ را به شکل یک مسئله ترارسانی منبع به هدف تلقی می کند و قوانین متناظرسازی را بین پیام های ورودی و پاسخ ها از یک مقدار حجیم داده های آموزشی یاد می گیرد. Ritter و همکارانش در سال 2011 مسئله تولید پاسخ را به شکل یک مسئله ترجمه ماشینی آماری SMT چارچوب بندی کرده است. Sordoni و همکارانش در سال 2015 سیستم Ritter و همکارانش را با نمره بندی مجدد خروجی های یک سیستم مکالمه مبتنی بر SMT عبارت دار به کمک یک مدل عصبی که زمینه قبلی را لحاظ می کند، بهبود داده اند. پیشرفت اخیر در مدلهای SEQ2SEQ الهام بخش اقدامات متعددی برای ایجاد سیستم های محاوره ای انتها به انتها بوده است که ابتدا یک کدگذاری کننده را برای متناظر سازی یک پیام به یک بردار توزیع شده نمایانگر اصول معنایی شناختی آن بکار می گیرد و یک پاسخی را از بردار پیام ایجاد می کند. Serban و همکارانش در سال 2016 یک مدل عصبی سلسله مراتبی را مطرح کردند که وابستگی ها را طی یک سابقه مکالمه بسط یافته کسب می کند. Li و همکارانش در سال 2016 اطلاع رسانی متقابل را میان پیام و پاسخ به شکل تابع عینی جایگزین برای کاهش نسبت پاسخ های عمومی ایجاد شده توسط سیستم های SEQ2SEQ مطرح کردند.

خط دیگر تحقیقات آماری بر ساخت سیستم های محاوره مبتنی بر وظیفه متمرکز بود تا کارهای خاص حوزه را حل نماید. تلاشها شامل مدلهای آماری نظیر فرایند تصمیم مارکوف MDP، مدلهای POMDP و مدلهایی که به لحاظ آماری قوانین ساخت را یاد می گیرند بوده است. این متون محاوره ای با اینحساب وسیعا به یادگیری پسخورد بکار بسته می شوند تا سیاستگزاریهای محاوره ای آموزش داده شود. ولیکن سیستم های محاوره ای RL مبتنی بر وظیفه اغلب متکی بر دقیقا پارامترهای محاوره ای محدود یا الگوهای دست ساز با سیگنالهای حالت، اقدام و پاداش

می باشد که توسط انسانها برای هر حوزه جدیدی طراحی شده است و این پارادیگم را برای بسط به شرح برنامه های حوزه باز دشوار می سازد.

همچنین کار قبلی درباره یادگیری پسخورد برای درک زبان (از جمله یادگیری از سیگنالهای پاداش تاخیری با انجام بازیهای مبتنی بر متن، اجرای دستورالعمل ها برای راهنمای ویندوز، یا درک دیالوگ هایی که جهات حرکت و جستجو را بدست می دهد) تحقیقی مرتبط می باشد.

هدف ما اضافه کردن SEQ2SEQ به پارادیگم های یادگیری پسخورد برای کسب مزیت های هر دو می باشد. با اینحساب بویژه از کار اخیر الهام گرفته ایم که تلاش دارد تا این پارادیگم ها را به یکدیگر اضافه کند از جمله کار Wen و همکارانش در سال 2016 (که در حال آموزش یک سیستم دیالوگ مبتنی بر وظیفه انتها به انتها که نمایشات ورودی را برای حل جفت ها براساس ترتیب ارزش در یک پایگاه داده بود) یا کار Su و همکارانش در سال 2016 که یادگیری پسخورد را با تولید عصبی طبق وظایف با کاربران واقعی ترکیب کردند و نشان دادند که یادگیری پسخورد باعث بهبود عملکرد مکالمه می شود.

3-یادگیری پسخورد برای مکالمه حوزه باز

در این بخش، ما به طور مفصل اجزای مدل RL مطرح شده را شرح می دهیم.

سیستم یادگیری شامل دو عامل می باشد. ما از p برای نشان دادن جملات ایجاد شده از اولین عامل و q برای نشان دادن جملات از دومین عامل استفاده می کنیم. دو عامل نوبت صحبت با یکدیگر را بدست می آورند. یک مکالمه می تواند به شکل یک توالی متناور از جملات ایجاد شده توسط دو عامل نمایش داده شود: $p_1, q_1, p_2, q_2, \cdots, p_i, q_i$. ما جملات ایجاد شده را به شکل اقداماتی می بینیم که طبق سیاست تعیین شده توسط یک مدل زبان شبکه عصبی متناوب کدگذاری کننده-کدگشایی کننده اتخاذ می شوند.

پارامترهای شبکه بهینه سازی می شوند تا پاداش آتی مورد انتظار را با استفاده از جستجوی سیاستگزاری طبق توضیحات بخش 4-3 به حداکثر برسانند. روشهای گرادیان سیاستگزاری برای سناریوی ما نسبت به یادگیری Q مناسب تر می باشند، چرا که می توانید RNN کدگذاری کننده-کدگشایی کننده را با استفاده از پارامترهای MLE که قبلا باعث ایجاد پاسخ های محتمل شده اند، قبل از تغییر هدف و تنظیم طبق یک سیاستگزاری که پاداش درازمدت را به حداکثر خود می رساند، راه اندازی نمایید. یادگیری Q از سوی دیگر مستقیما باعث تخمین پاداش

مورد انتظار آتی هر اقدام می شود که می تواند از هدف MLE بنا به ترتیب بزرگی متفاوت باشد، و با اینحساب باعث شود که پارامترهای MLE برای راه اندازی نامناسب شوند. مولفه ها (حالات، پاداش و غیره) مسئله تصمیم متوالی ما در قسمتهای فرعی ذیل خلاصه سازی شده اند.

3-1-اقدام

یک اقدام a بیان مکالمه برای ایجاد است. فضای اقدام نامحدود است چرا که توالی های با طول اختیاری را می توان ایجاد کرد.

3-2-حالت

یک حالت با دو نوبت دیالوگ $[p_i, q_i]$ قبلی نمایش داده می شود. سابقه مکالمه باز به یک نمایش برداری تغییر شکل می یابد که با تغذیه غلظت p_i و q_i به یک مدل کدگذاری کننده LSTM طبق توضیح Li و همکارانش می باشد.

3-3-سیاستگزاری

یک سیاستگزاری شکلی از یک کدگذاری کننده-کدگشاینده LSTM را (یعنی $p_{RL}(p_{i+1}|p_i, q_i)$) به خود می گیرد و با پارامترهایش تعریف می شود. توجه داشته باشید که ما از یک نمایش فاجعه آمیز سیاستگزاری استفاده می کنیم (یعنی توزیع احتمالات روی اقدامات با درنظرگیری حالات). یک سیاستگزاری قاطع منجر به یک هدف گسسته می شود که بهینه سازی را با استفاده از روشهای مبتنی بر گرادیان مشکل می سازد.

3-4-جایزه

r نشانه پاداش بدست آمده برای هر اقدام می باشد. در این قسمت کوچک، ما درباره عوامل اصلی بحث می کنیم که در موفقیت یک مکالمه نقش دارد و توضیح می دهد که چگونه تخمین ها در این عوامل می تواند در عملکردهای پاداش قابل محاسبه عملی گردد.

سهولت پاسخ دهی. یک نوبت مکالمه تولید شده توسط یک ماشین باید به سهولت پاسخ داده شود. این جنبه از یک نوبت مکالمه به عملکرد اینده نگرانه اش مربوط می شود: یعنی محدودیت هایی که یک نوبت مکالمه برای نوبت بعدی به همراه دارد. ما مطرح کرده ایم که سهولت پاسخ دهی به یک نوبت مکالمه ایجاد شده با استفاده از

احتمال لگاریتمی منفی پاسخ دهی به آن بیان با یک پاسخ کسل کننده اندازه گیری شود. ما به طور دستی یک فهرست از پاسخ های کسل کننده S را شامل 8 نوبت مکالمه مانند «نمی دونم درباره چی صحبت می کنی.»، «هیچ ایده ای ندارم» و غیره ساخته ایم که ما و سایرین دریافته ایم که خیلی زیاد در مدلهای SEQ2SEQ مکالمات رخ می دهد. عملکرد پاداش به ترتیب ذیل معین می شود:

$$r_1 = -\frac{1}{N_{\mathbb{S}}} \sum_{s \in \mathbb{S}} \frac{1}{N_s} \log p_{\text{seq2seq}}(s|a) \qquad (1)$$

که در آن $N_{\mathbb{S}}$ نمایانگر اصلی بودن $N_{\mathbb{S}}$ می باشد و N_s نشانگر تعداد نمونه ها در پاسخ کسل کننده s می باشد. هر چند البته راههای بیشتری برای ایجاد پاسخ کسل کننده نسبت به آنی که فهرست بتواند تحت پوشش قرار دهد، وجود دارد، بسیاری از این پاسخ ها احیانا در نواحی مشابه در فضای برداری محاسبه شده توسط مدل قرار می گیرد. با اینحساب، یک سیستم که به احتمال کمتری بیان های این فهرست را ایجاد می کند، همچنین به احتمال کمتری سایر پاسخ های کسل کننده را ایجاد می کند.

p_{seq2seq} نمایانگر خروجی محتمل توسط مدلهای SEQ2SEQ می باشد. لازم به ذکر است که p_{seq2seq} متفاوت از تابع سیاستگزاری تصادفی $p_{RL}(p_{i+1}|p_i, q_i)$ می باشد، چون اولی براساس هدف MLE در مدل SEQ2SEQ یاد گرفته می شود درحالیکه دومی سیاستگزاری بهینه سازی شده برای پاداش درازمدت در آینده در شرایط RL می باشد. r_1 باز طبق طول هدف S مقیاس بندی می شود.

جریان اطلاعات. ما می خواهیم که هر عامل در اطلاعات جدید در هر نوبت مکالمه برای حفظ تداوم مکالمه و اجتناب از توالی تکراری اش همکاری نماید. ما از اینرو جریمه سازی تشابه معنایی را بین نوبت های متوالی مکالمه از جانب همان عامل مطرح کرده ایم. اجازه دهید که h_{p_i} و $h_{p_{i+1}}$ نشانگر نمایش های بدست آمده از کدگذاری کننده برای دو نوبت متوالی مکالمه p_i و p_{i+1} باشد. پاداش با لگاریتم منفی کسینوس شباهت بین آنها معین می شود:

$$r_2 = -\log\cos(h_{p_i}, h_{p_{i+1}}) = -\log\cos\frac{h_{p_i}\cdot h_{p_{i+1}}}{\|h_{p_i}\|\|h_{p_{i+1}}\|} \quad (2)$$

انسجام معنایی. همچنین باید کفایت پاسخ ها را برای اجتناب از موقعیت ها اندازه گیری نمایید که در آن پاسخ های ایجاد شده به شدت پاداش داده می شوند ولیکن گرامری نیستند و منسجم هم نیستند. ما از اینرو اطلاع رسانی متقابل را میان اقدام a و نوبت های مکالمه قبلی در سابقه مکالمه درنظر گرفته ایم تا تضمین کنیم که پاسخ های ایجاد شده منسجم و مناسب می باشد:

$$r_3 = \frac{1}{N_a}\log p_{\text{seq2seq}}(a|q_i, p_i) + \frac{1}{N_{q_i}}\log p^{\text{backward}}_{\text{seq2seq}}(q_i|a) \quad (3)$$

$p_{\text{seq2seq}}(a|p_i, q_i)$ نشانگر احتمال ایجاد پاسخ a با درنظرگیری بیان محاوره ای قبلی $[p_i, q_i]$ می باشد. $p^{\text{backward}}_{\text{seq2seq}}(q_i|a)$ نشانگر احتمال رو به عقب ایجاد بیان محاوره ای قبلی q_i براساس پاسخ a می باشد. $p^{\text{backward}}_{\text{seq2seq}}$ به همان شیوه مشابه مدلهای استاندارد SEQ2SEQ با منابع و اهداف پایاپای آموزش می بیند. باز، برای کنترل اثر طول هدف، هم $\log p_{\text{seq2seq}}(a|q_i, p_i)$ و هم $\log p^{\text{backward}}_{\text{seq2seq}}(q_i|a)$ براساس طول اهداف مقیاس بندی می شوند.

پاداش نهایی برای اقدام a یک حاصل جمع سنجیده شده پاداش های بحث شده قبلی می باشد:

$$r(a, [p_i, q_i]) = \lambda_1 r_1 + \lambda_2 r_2 + \lambda_3 r_3 \quad (4)$$

که در آن $\lambda_1 + \lambda_2 + \lambda_3 = 1$ می باشد. ما $\lambda_1 = 0.25, \lambda_2 = 0.25$ و $\lambda_3 = 0.5$ را تعیین کرده ایم. یک پاداش بعد از اینکه عامل به انتهای هر جمله رسید، مشاهده می شود.

4-شبیه سازی

ایده اصلی پشت روش ما همان شبیه سازی فرایند دو عامل مجازی است که نوبت صحبت را با یکدیگر گرفته اند که از طریق آن می توانیم فضای حالت-اقدام را بررسی نماییم و یک سیاستگزاری $p_{RL}(p_{i+1}|p_i, q_i)$ را یاد بگیریم که منجر به پاداش مورد انتظار بهینه می شود. ما یک راهکار سبک AlphaGo را با راه اندازی سیستم RL با استفاده از یک سیاستگزاری ایجاد پاسخ عمومی اتخاذ کرده ایم که از یک شرایط کاملا تحت نظارت یاد گرفته شده است.

4-1-یادگیری تحت نظارت

برای اولین مرحله یادگیری، ما طبق کار قبلی پیشگویی یک توالی هدف ایجاد شده را با درنظرگیری سابقه محاوره با استفاده از مدل تحت نظارت SEQ2SEQ ساخته ایم. نتایج حاصل از مدلهای تحت نظارت بعدا برای راه اندازی مدل استفاده خواهند شد.

ما یک مدل SEQ2SEQ را با توجه به پایگاه داده OpenSubtitles آموزش داده ایم که شامل تقریبا 80 میلیون جفت منبع-هدف بوده اند. ما هر نوبت را در پایگاه داده ها به شکل یک هدف تلقی کرده و تسلسل دو جمله قبلی را به شکل ورودی های منبع درنظر گرفته ایم.

4-2-اطلاع رسانی متقابل

نمونه های حاصل از مدل های SEQ2SEQ اغلب اوقات کسل کننده و عمومی می باشند نظیر «نمی دونم». با اینحساب، ما نمی خواهیم مدل سیاستگزاری را با استفاده از مدلهای SEQ2SEQ از قبل آموزش دیده راه اندازی کنیم چرا که این امر منجر به فقدان تنوع در تجربیات مدلهای RL خواهد شد. Li و همکارانش در 2016 نشان دادند که مدلسازی اطلاع رسانی متقابل بین منابع و اهداف به طور معنی داری شانس ایجاد پاسخ های کسل کننده را کاهش داده و باعث بهبود کیفیت پاسخ دهی می شود. ما اکنون نشان داده ایم که چگونه یک مدل کدگذاری کننده-کدگشاینده را می توانیم بدست آوریم که باعث ایجاد پاسخ های اطلاع رسانی متقابل حداکثر می شود.

همانگونه که در مقاله Li و همکارانش در سال 2016 نشان داده شده است، کدگشایی مستقیم از معادله 3 غیرعملی است چرا که عبارت دوم نیاز دارد که جمله هدف کاملا ایجاد گردد. در این کار که از کار اخیر درباره

یادگیری سطح متوالی الهام گرفته شده است، ما مسئله ایجاد پاسخ اطلاع رسانی متقابل ماکزیمم را به عنوان مسئله یادگیری پسخورد تلقی می کنیم که در آن یک پاداش ارزش اطلاع رسانی متقابل زمانی مشاهده می شود که مدل به پایان یک توالی رسیده باشد.

مشابه Ranzato و همکارانش در سال 2015، ما از روشهای گرادیان سیاستگزاری برای بهینه سازی استفاده کردیم. ما مدل سیاستگزاری p_{RL} را با استفاده از یک مدل $p_{\text{SEQ2SEQ}}(a|p_i, q_i)$ از قبل آموزش دیده راه اندازی کرده ایم. با درنظرگیری یک منبع ورودی $[p_i, q_i]$، ما یک فهرست کاندیدای $A = \{\hat{a}|\hat{a} \sim p_{RL}\}$ را ایجاد می کنیم. برای هر کاندیدای ایجاد شده $\hat{a}$، ما امتیاز اطلاع رسانی متقابل $m(\hat{a}, [p_i, q_i])$ را از $p_{\text{SEQ2SEQ}}(a|p_i, q_i)$ و $p^{\text{backward}}_{\text{SEQ2SEQ}}(q_i|a)$ از قبل آموزش دیده بدست خواهیم آورد. این امتیاز اطلاع رسانی متقابل به عنوان یک پاداش استفاده خواهد شد و به مدل کدگذاری کننده-کدگشاینده انتشار رو به عقب خواهد داشت،و آنرا برای ایجاد عواقبی با پاداش های بالاتر سازگار خواهد کرد. ما خوانندگان را به مطالعه Zaremba & Sutskever در سال 2015 و Williams در سال 1992 برای جزئیات بیشتر ارجاع می دهیم. پاداش مورد انتظار برای یک توالی به ترتیب ذیل معین می شود:

$$J(\theta) = \mathbb{E}[m(\hat{a}, [p_i, q_i])] \qquad (5)$$

گرادیان با استفاده از حقه نسبت احتمالات تخمین زده می شود:

$$\nabla J(\theta) = m(\hat{a}, [p_i, q_i]) \nabla \log p_{RL}(\hat{a}|[p_i, q_i]) \qquad (6)$$

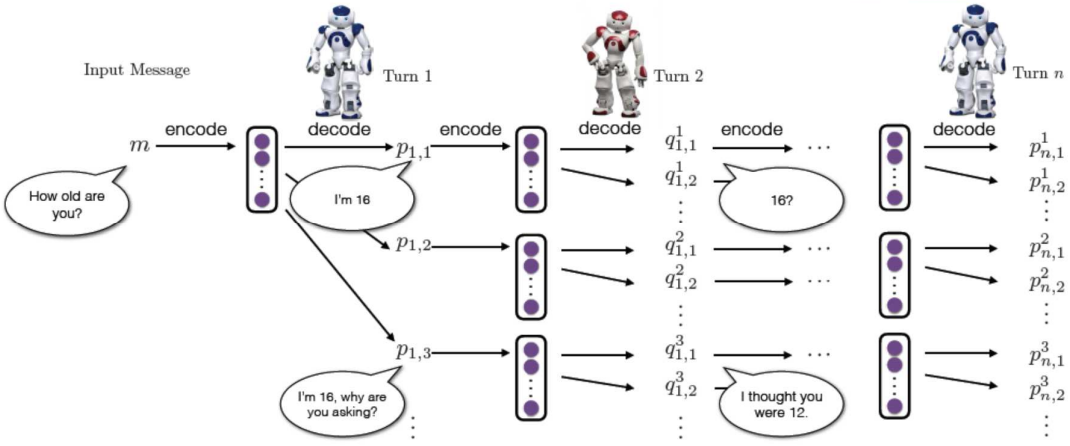
ما پارامترها را در مدل کدگشاینده-کدگذاری کننده با ستفاده از نزول گرادیان تصادفی روزآمدسازی کرده ایم. یک راهکار یادگیری برنامه درسی همانند مطالعه Ranzato و همکارانش در سال 2015 اتخاذ گردیده است به نحوی که برای هر توالی به طول T ما از کمبود MLE برای اولین L نمونه ها و الگوریتم پسخورد برای بقیه نمونه های T-L استفاده کرده ایم. ما مقدار L را به صفر تقویت سازی کرده ایم. یک راهکار خط پایه برای کاهش واریانس یادگیری بکار گرفته شده است: یک مدل خنثی دیگر به عنوان ورودی هدف ایجاد شده و منبع اولیه را گرفته و

یک مقدار پایه را به عنوان خروجی گرفته است که مشابه با راهکار اتخاذ شده توسط Zaremba & Sutskever می باشد. گرادیان نهایی از اینرو برابر است با:

$$\nabla J(\theta) = \nabla \log p_{RL}(\hat{a}|[p_i, q_i])[m(\hat{a}, [p_i, q_i]) - b] \quad (7)$$

3-4-شبیه سازی دیالوگ بین دو عامل

ما مکالمات را میان دو عامل مجازی شبیه سازی کرده ایم و به آنها نوبت صحبت کردن به یکدیگر دادیم. شبیه سازی به ترتیب ذیل پیش رفته است: در مرحله اولیه، یک پیام از جانب مجموعه آموزشی اولین عامل را تغذیه می کند. عامل آن پیام ورودی را به یک نمایش برداری کدگذاری کرده و شروع به کدگشایی می کند تا یک نتیجه خروجی پاسخ را ایجاد کند. با لحاظ خروجی فوری از اولین عامل با سابقه مکالمه، دومین عامل حالت را با کدگذاری سابقه محاوره به یک نمایش روزآمدسازی می کند و از کدگشاینده RNN برای ایجاد پاسخ ها استفاده می نماید، که به طور متوالی به اولین عامل تغذیه می شود و فرایند تکرار می شود.



شکل 1-شبیه سازی مکالمه بین دو عامل

بهینه سازی- ما مدل سیاستگزاری p_{RL} را با پارامترهایی از مدل اطلاع رسانی متقابل که در بخش قبلی توضیح داده شده، راه اندازی می نماییم. سپس از روشهای گرادیان سیاستگزاری برای یافتن پارامترهایی استفاده می کنیم که منجر به یک پاداش مورد انتظار بزرگتر می شود. هدف از به حداکثررسانی همان پاداش آتی مورد انتظار می باشد:

$$J_{RL}(\theta) = \mathbb{E}_{p_{RL}(a_{1:T})}[\sum_{i=1}^{i=T} R(a_i, [p_i, q_i])] \quad (8)$$

که در آن $R(a_i, [p_i, q_i])$ نشانگر پاداش ناشی از عملکرد a_i می باشد. ما از حقه نسبت احتمالات برای روزآمدسازی های گرادیان استفاده می کنیم:

$$\nabla J_{RL}(\theta) \approx \sum_i \nabla \log p(a_i|p_i, q_i) \sum_{i=1}^{i=T} R(a_i, [p_i, q_i]) \quad (9)$$

ما خوانندگان را به مطالعه Williams در سال 1992 و Glynn در سال 1990 برای جزئیات بیشتر ارجاع می دهیم.

4-4-یادگیری برنامه درسی

یک راهکار یادگیری برنامه درسی باز بکار گرفته می شود که در آن ما شروع به شبیه سازی مکالمه برای دو نوبت مکالمه می نماییم، و به تدریج تعداد نوبت های مکالمه شبیه سازی شده را افزایش می دهیم. ما 5 نوبت مکالمه را ایجاد می کند، حین اینکه تعداد کاندیداها برای بررسی به طور نمایی از لحاظ اندازه فهرست کاندیدا رشد می کند. پنج پاسخ کاندیدا در هر مرحله شبیه سازی ایجاد می شود.

5-نتایج آزمایشی

در این بخش، ما نتایج آزمایشی را در کنار تحلیل کیفی ما سیستم های ایجاد محاوره را هم با استفاده از قضاوت های انسانی و هم دو اصول سنجش خودکار ارزیابی کرده ایم: طول مکالمه (تعداد نوبت های مکالمه در جلسه کامل) و تنوع.

1-5-پایگاه داده ها

شبیه سازی مکالمه نیاز به ورودی های اولیه با کیفیت بالا برای تغذیه عامل دارد. برای مثال، یک ورودی اولیه «چرا؟» مطلوب نمی باشد چرا که مشخص نیست که چگونه دیالوگ می تواند پیش برود. ما یک زیرمجموعه از ده میلیون پیام را از پایگاه داده OpenSubtitles برگرفته ایم و تعداد 0.8 میلیون توالی را با پایین ترین احتمال

ایجاد پاسخ «نمی دونم درباره چی حرف می زنی؟» را برای تضمین اینکه به ورودی های اولیه راحت پاسخ داده می شود.

2-5-ارزیابی خودکار

ارزیابی سیستم های محاوره مشکل است اصول سنجشی مانند BLEU و پیچیدگی وسیعا برای ارزیابی کیفیت محاوره استفاده شده است، ولیکن وسیعا مناظره شده که این اصول سنجش خودکار به چه خوبی با کیفیت پاسخ حقیقی همبستگی دارند. چون هدف از سیستم مطرح شده پیشگویی بالاترین پاسخ احتمالات نمی باشد، بلکه در عوض موفقیت طولانی مدت محاوره است، ما BLEU یا پیچیدگی را برای ارزیابی بکار نمی بریم.[۲]

مدل	شماره نوبت های شبیه سازی شده
SEQ2SEQ	2.68
اطلاع رسانی	3.40
متقابل	4.48
RL	

جدول 2-متوسط تعداد نوبت های شبیه سازی شده از مدلهای استاندارد SEQ2SEQ2، مدل اطلاعات و مدل RL مطرح شده.

طول مدت مکالمه-اصول سنجش اولیه که مطرح کرده ایم طول مکالمه شبیه سازی شده است. ما می گوییم یک مکالمه زمانی پایان می یابد که یکی از عوامل شروع به ایجاد پاسخ های کسل کننده ای نظیر «نمی دونم»[۳] کنند یا دو بیان متوالی از همان کاربر به شدت همپوشانی داشته باشد.[۴]

مجموعه تست شامل هزار پیام ورودی می باشد. برای کاهش ریسک مکالمات دایره ای، ما تعداد برگشت های شبیه سازی شده را به کمتر از 8 محدود کرده ایم. نتایج در جدول 2 نشان داده شده است. همانطور که می توان

---

[2] ما دریافته ایم که مدل RL در زمینه نمره BLEU بدتر عمل کرده است. در یک نمونه تصادفی متشکل از 2500 جفت مکالمه، امتیازات BLEU مرجع منفرد برای مدلهای RL، مدلهای اطلاع رسانی متقابل و مدلهای ساده SEQ2SEQ به ترتیب برابر است با 1.28 و 1.44 و 1.17. BLEU به شدت با پیچیدگی در کارهای ساخت همبستگی داشته است. چون مدل RL براساس پاداش آتی به جای MLE آموزش دیده است، تعجبی ندارد که مدلهای مبتنی بر RL به امتیاز BLEU کمتری دست می یابند.

[3] ما از یک روش تطابق قانون ساده با یک فهرست از 8 عبارت استفاده کرده ایم که به شکل پاسخ های کسل کننده شمارش شده اند. هرچند این امر می تواند منجر به هم کاذب مثبت و هم کاذب منفی شود، به خوبی در عمل کار می کند.

[4] دو بیان به نظر تکراری می آیند اگر بیش از 80 درصد کلماتشان مشترک باشند.

دید، استفاده از اطلاع رسانی متقابل منجربه مکالمات پایدارتری بین دو عامل شده است. مدل RL مطرح شده ابتدا براساس هدف اطلاع رسانی متقابل آموزش دیده و با اینحساب از آن علاوه بر مدل RL سود می برد. ما مشاهده کرده ایم که مدل RL با شبیه سازی محاوره ای به بهترین نمره ارزیابی دست می یابد.

تنوع-ما درجه تنوع را با محاسبه تعداد تک گرام و دابل گرام های مجزا در پاسخ های ایجاد شده گزارش داده ایم. ارزش بنا به تعداد کل نمونه های ایجاد شده برای اجتناب از جملات طولانی دلخواه بنا به شرح Li و همکارانش در سال 2016 مقیاس بندی شده است. اصول سنجش منتج با اینحساب یک نسبت نمونه نوعی برای تک گرام ها و دابل گرام ها می باشد.

برای هر دو مدل استاندارد SEQ2SEQ و مدل RL مطرح شده، ما از جستجوی پرتویی با یک اندازه پرتوی 10 برای ایجاد یک پاسخ به یک پیام ورودی معین استفاده می کنیم. برای مدل اطلاع رسانی متقابل، ما ابتدا فهرستی از n بهترین را با استفاده از $p_{\text{SEQ2SEQ}}(t|s)$ ایجاد می کنیم و بعد به طور خطی آنها را با استفاده از $p_{\text{SEQ2SEQ}}(s|t)$ رتبه بندی مجدد می کنیم. نتایج در جدول 4 ارائه شده اند. ما دریافته ایم که مدل RL مطرح شده نتایج متنوع تری را ایجاد کرده اند وقتی که طبق هم مدل ساده SEQ2SEQ و هم مدل اطلاع رسانی متقابل مقایسه می شوند.

مدل	تکگرام	دابلگرام
SEQ2SEQ	0.0062	0.015
اطلاع رسانی متقابل	0.011	0.031
RL	0.017	0.041

جدول 4-امتیازات تنوع (نسبت نمونه به نوع) برای مدل استاندارد SEQ2SEQ، مدل اطلاع رسانی متقابل و مدل RL مطرح شده

ارزیابی انسانی –ما سه شرایط را برای ارزیابی انسانی بررسی کرده ایم: اولین شرایط مشابه با آنی است که در مطالعه Li و همکارانش در سال 2016 توضیح داده شده است، که در آن ما قضاوت کننده های منبع توده ای را برای ارزیابی یک نمونه تصادفی از 500 گزینه را بکار بسته ایم. ما هم یک پیام ورودی و هم خروجی های ایجاد شده را به سه قضاوت کننده ارائه کرده ایم و از آنها خواسته ایم که تصمیم بگیرند کدام یک از دو خروجی نتیجه

بهتر است (که به صورت کیفیت عمومی تک نوبتی نشان داده شده است). بندها مجاز است. رشته های یکسان امتیاز یکسان دارند. ما بهبود حاصله توسط مدل RL را طی مدل اطلاع رسانی متقابل با میانگین تفاوت در امتیازات میان مدل ها اندازه گیری کرده ایم.

برای شرایط دوم، به قضاوت کنندگان مجددا پیام های ورودی و خروجی های سیستم نشان داده شد ولیکن از آنها خواسته شده که تصمیم بگیرند کدامیک از دو نتیجه خروجی پاسخ دهی آسان تری دارد (که با سهولت تک نوبتی برای پاسخ دهی نشان داده شده است). باز ما یک نمونه تصادفی متشکل از 500 گزینه را ارزیابی کرده ایم که هر یک به 3 قضاوت کننده اختصاص داده شده اند.

برای سومین شرایط، به قضاوت کنندگان مکالمات شبیه سازی شده ای بین دو عامل نشان داده میشود (که به شکل کیفیت عمومی چندنوبتی نشان داده می شود). هر مکالمه شامل 5 نوبت می باشد. ما به ارزیابی 200 مکالمه شبیه سازی شده می پردازیم، که هر یک به 3 قضاوت کننده اختصاص دارد، که از آنها خواسته شده تا تصمیم بگیرند کدامیک از مکالمات شبیه سازی شده از کیفیت بالاتری برخوردارند.

شرایط	برنده RL	بازنده RL	بند
کیفیت عمومی تک نوبتی	0.40	0.36	0.24
سهولت تک نوبتی برای پاسخ دهی	0.52	0.23	0.25
کیفیت عمومی چندنوبتی	0.72	0.12	0.16

جدول 5-RL روی سیستم اطلاع رسانی متقابل مبتنی بر قضاوت های انسانی جفتی بدست می آید.

نتایج برای ارزیابی انسانی در جدول 5 نشان داده شده است. سیستم RL مطرح شده یک تقویت اساسی را در کیفیت پاسخ تک نوبتی (برندگی در 40 درصد مواقع و بازندگی در 36 درصد مواقع) را ایجاد نکرده است.

پیام ورودی	مدل اطلاع رسانی متقابل	RL
حالت چطوره؟ نام کاملت چیست؟ نمی خواهم امشب برم خونه.	شانزده سالمه. هیچ نظری ندارم. واقعا؟	شانزده سالمه، چرا می پرسی؟ اسم تو چیه؟ چرا؟ نمی خوام به احساساتت لطمه بزنم.

ده ثانیه بلی. کار زیادی برای انجام دادن در اینجا داریم. بلی. من فوتبال دوست دارم. درباره چی؟ فکر نکنم ایده خوبی برای گفتن این مطلب باشه.	نمی دونم درباره چی حرف می زنی. نه به قدر کافی طولانی. متاسفم، آقا. البته! نه انجام ندادم. منظورم اینه که مرد خوبیه. چون نمی خوام صدمه ببینی	هیچ احساسی به من داری؟ چه مدت دیگه اینجایی؟ باید شروع کنیم؟ فوتبال بازی می کنی؟ باید با اون حرف بزنیم. چرا تا حالا درباره اش نگفتی؟

جدول 3-پاسخ های نمونه برداری شده ایجاد شده از مدلهای اطلاع رسانی متقابل و مدل RL مطرح شده.

این امر همتراز با انتظارات ما می باشد، همانگونه که مدل RL بهینه سازی نشده تا بیان بعدی را بگوید ولیکن در عوض پاداش درازمدتی را افزایش می دهد. سیستم RL پاسخ هایی را ایجاد می کند که به طور معنی داری پاسخ دهی سهل تری نسبت به سیستم اطلاع رسانی متقابل دارد، همانگونه که توسط سهولت تک نوبتی به شرایط پاسخ دهی (با برندگی 52 درصد زمان و از دست دادن 23 درصد زمان) و نیز با دیالوگ های چند نوبتی با کیفیت بالاتر معنی دار بنا به شرایط کیفیت عمومی چند نوبتی (با برندگی در 72 درصد مواقع) نشان داده شده است.

تحلیل کیفی و بحث-ما یک نمونه تصادفی از پاسخ های ایجاد شده را در جدول 3 نشان داده ایم و مکالمات را در جدول 1 در آغاز مقاله شبیه سازی کرده ایم. از روی جدول 3، می توانیم ببینیم که عامل مبتنی بر RL در واقع پاسخ های تعاملی تری را نسبت به حالت های پایه دیگر ایجاد می کند. همچنین دریافته ایم که مدل RL دارای تمایل برای به اتمام رسانیدن یک جمله با سوال دیگر و تحویل دادن مکالمه به کاربر است. از روی جدول 1، مشاهده می کنیم که مدل RL مدیریت می کند تا مکالمات تعاملی تر و پایدارتری را نسبت به مدل اطلاع رسانی متقابل ایجاد نماید.

طی تحلیل خطا، دریافتیم که هر چند ما بیان های تکراری را در نوبت های متوالی جریمه سازی کرده ایم، مکالمه گاهی اوقات به چرخه ای با طول بیش از یک بنا به جدول 6 وارد می شود. این امر می تواند به مقدار محدود سابقه مکالماتی که درنظر گرفته ایم نسبت داده شود. مسئله دیگر مشاهده شده این است که این مدل گاهی اوقات موضوع نامرتبطی را طی مکالمه شروع می کند. یک تبادلی میان مرتبط بودن و تکرار کمتر بنا به عملکرد پاداش که در معادله 4 تعریف کرده ایم، وجود دارد.

البته مسئله بنیانی این است که عملکرد پاداش تعریف شده دستی نمی تواند احتمالا جنبه های قاطعی را بپوشاند که یک مکالمه ایده آل را تعریف می کند.

الف-اسمت چیه؟
ب-دنیل.
الف-چند سالته؟
ب-دوازده سال. اسم تو چیه؟
الف-دنیل.
ب-چند سالته؟
الف-دوازده سال. اسم تو چیه؟
ب-دنیل.
الف-چند سالته؟
ب-....

جدول 6-یک دیالوگ شبیه سازی شده با یک چرخه طولانی تر از یک

در حالیکه پاداش های مکاشفه ای که تعریف کرده ایم مسئول محاسبه خودکار می باشند، و برخی جنبه هایی را از آنچه یک مکالمه خوب را می سازد، کسب می کند یعنی به طور ایده آل سیستم در عوض پاداش های واقعی را از انسانها دریافت می کند. مسئله دیگر با مدل کنونی این است که ما می توانیم فقط بررسی یک تعداد اندکی از کاندیداها را انجام دهیم و نوبت هایی را شبیه سازی کنیم چون تعداد موارد درنظر گرفته شده به طور نمایی رشد می کند.

6-نتیجه گیری

ما یک چارچوب یادگیری پسخورد را برای ایجاد پاسخ عصبی با شبیه سازی مکالمات بین دو عامل با لحاظ کردن نقاط قوت سیستم های عصبی SEQ2SEQ و یادگیری پسخورد برای مکالمه وارد کرده ایم. نظیر مدلهای عصبی قبلی تر SEQ2SEQ، چارچوب ما مدلهای انشایی معنای یک نوبت مکالمه را کسب کرده و پاسخ های مناسب معنایی ایجاد می کند. نظیر سیستم های مکالمه یادگیری پسخورد، چارچوب ما قادر است بیان هایی را ایجاد کند که پاداش اتی را بهینه سازی می کند و به طور موفقیت امیزی خصوصیات جهانی یک مکالمه خوب را کسب می کند. علی رغم این حقیقت که مدل ما از مکاشفه خیلی ساده قابل اجرا برای کسب این خصوصیات جهانی استفاده می کند، این چارچوب پاسخ های تعاملی متنوع تری را نسبت به شکوفایی یک مکالمه پایدارتر ایجاد می کند.

References

V. M. Aleksandrov, V. I. Sysoyev, and V. V. Shemeneva. 1968. Stochastic optimization. *Engineering Cybernetics*, 5:11–16.

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9:1–26.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. of ICLR*.

Rafael E Banchs and Haizhou Li. 2012. IRIS: a chat-oriented dialogue system based on the vector space model. In *Proceedings of the ACL 2012 System Demonstrations*, pages 37–42.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM.

SRK Branavan, David Silver, and Regina Barzilay. 2011. Learning to win by reading manuals in a monte-carlo framework. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 268–277.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proc. of ACL-IJCNLP*, pages 445–450, Beijing, China, July.

Milica Gašic, Catherine Breslin, Matthew Henderson, Dongho Kim, Martin Szummer, Blaise Thomson, Pirros Tsiakoulis, and Steve Young. 2013a. Pomdp-based dialogue manager adaptation to extended domains. In *Proceedings of SIGDIAL*.

Milica Gasic, Catherine Breslin, Mike Henderson, Dongkyu Kim, Martin Szummer, Blaise Thomson, Pirros Tsiakoulis, and Steve Young. 2013b. On-line policy optimisation of bayesian spoken dialogue systems via human interaction. In *Proceedings of ICASSP 2013*, pages 8367–8371. IEEE.

Milica Gašic, Dongho Kim, Pirros Tsiakoulis, Catherine Breslin, Matthew Henderson, Martin Szummer, Blaise Thomson, and Steve Young. 2014. Incremental on-line adaptation of pomdp-based dialogue managers to extended domains. *In Proceedings on InterSpeech*.

Peter W Glynn. 1990. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33(10):75–84.

Ji He, Jianshu Chen, Xiaodong He, Jianfeng Gao, Lihong Li, Li Deng, and Mari Ostendorf. 2016. Deep reinforcement learning with a natural language action space. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1621–1630, Berlin, Germany, August.

Esther Levin, Roberto Pieraccini, and Wieland Eckert. 1997. Learning dialogue strategies within the markov decision process framework. In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 72–79. IEEE.

Esther Levin, Roberto Pieraccini, and Wieland Eckert. 2000. A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on Speech and Audio Processing*, 8(1):11–23.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proc. of NAACL-HLT*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany, August.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Yi Luan, Yangfeng Ji, and Mari Ostendorf. 2016. LSTM based conversation models. *arXiv preprint arXiv:1603.09457*.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing Atari with deep reinforcement learning. *NIPS Deep Learning Workshop*.

Karthik Narasimhan, Tejas Kulkarni, and Regina Barzilay. 2015. Language understanding for text-based games using deep reinforcement learning. *arXiv preprint arXiv:1506.08941*.

Lasguido Nio, Sakriani Sakti, Graham Neubig, Tomoki Toda, Mirna Adriani, and Satoshi Nakamura. 2014. Developing non-goal dialog system based on examples of drama television. In *Natural Interaction with Robots, Knowbots and Smartphones*, pages 355–361. Springer.

Alice H Oh and Alexander I Rudnicky. 2000. Stochastic language generation for spoken dialogue systems. In *Proceedings of the 2000 ANLP/NAACL Workshop on Conversational systems-Volume 3*, pages 27–32.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318.

Roberto Pieraccini, David Suendermann, Krishna Dayanidhi, and Jackson Liscombe. 2009. Are we there yet? Research in commercial spoken dialog systems. In *Text, Speech and Dialogue*, pages 3–13. Springer.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Adwait Ratnaparkhi. 2002. Trainable approaches to surface natural language generation and their application to conversational dialog systems. *Computer Speech & Language*, 16(3):435–455.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of EMNLP 2011*, pages 583–593.

Jost Schatzmann, Karl Weilhammer, Matt Stuttle, and Steve Young. 2006. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The knowledge engineering review*, 21(02):97–126.

Emanuel A. Schegloff and Harvey Sacks. 1973. Opening up closings. *Semiotica*, 8(4):289–327.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of AAAI*, February.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of ACL-IJCNLP*, pages 1577–1586.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

Satinder P Singh, Michael J Kearns, Diane J Litman, and Marilyn A Walker. 1999. Reinforcement learning for spoken dialogue systems. In *Nips*, pages 956–962.

Satinder Singh, Michael Kearns, Diane J Litman, Marilyn A Walker, et al. 2000. Empirical evaluation of a reinforcement learning spoken dialogue system. In *AAAI/IAAI*, pages 645–651.

Satinder Singh, Diane Litman, Michael Kearns, and Marilyn Walker. 2002. Optimizing dialogue management with reinforcement learning: Experiments with the njfun system. *Journal of Artificial Intelligence Research*, pages 105–133.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Meg Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of NAACL-HLT*.

Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Continuously learning neural dialogue management. *arxiv*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Richard S Sutton, David A McAllester, Satinder P Singh, Yishay Mansour, et al. 1999. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, volume 99, pages 1057–1063.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *Proceedings of ICML Deep Learning Workshop*.

Adam Vogel and Dan Jurafsky. 2010. Learning to follow navigational directions. In *Proceedings of ACL 2010*, pages 806–814.

Marilyn A Walker, Rashmi Prasad, and Amanda Stent. 2003. A trainable generator for recommendations in multimodal dialog. In *Proceeedings of INTERSPEECH 2003*.

Marilyn A. Walker. 2000. An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artificial Intelligence Research*, pages 387–416.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of EMNLP*, pages 1711–1721, Lisbon, Portugal.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, and Steve Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, and Xiaolong Wang. 2016. Incorporating loose-structured knowledge into LSTM with recall gate for conversation modeling. *arXiv preprint arXiv:1605.05110*.

Kaisheng Yao, Geoffrey Zweig, and Baolin Peng. 2015. Attention with intention for a neural network conversation model. In *NIPS workshop on Machine Learning for Spoken Language Understanding and Interaction*.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.

Steve Young, Milica Gasic, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Wojciech Zaremba and Ilya Sutskever. 2015. Reinforcement learning neural Turing machines. *arXiv preprint arXiv:1505.00521*.