جمع‌بندی و مرور ادبیات پژوهشی

Reinforcement using Facebook on Text Ad Generative Improving Learning

نام و نام خانوادگی: هادی فتحی‌پور
شماره دانشجویی: ۴۰۴۱۱۳۳۴
درس پردازش زبان طبیعی

۱ مقدمه و شرح مسئله

با پیشرفت مدل‌های زبانی بزرگ (Large Language Models)، کاربرد آن‌ها در مسائل واقعی پردازش زبان طبیعی به‌طور چشم‌گیری افزایش یافته است. یکی از مهم‌ترین این کاربردها، تولید متن در سیستم‌های تبلیغات آنلاین است؛ جایی که کیفیت متن نه‌تنها از نظر زبانی، بلکه از نظر تأثیرگذاری بر رفتار کاربران واقعی اهمیت دارد.

در سیستم‌های تبلیغاتی، معیارهایی مانند نرخ کلیک (Click-Through Rate) یا CTR به‌عنوان شاخص‌های عینی برای سنجش کیفیت متن تبلیغاتی استفاده می‌شوند. با این حال، روش‌های متداول آموزش مدل‌های زبانی، مانند Supervised Fine-Tuning، معمولاً بر تقلید از داده‌های متنی موجود تمرکز دارند و به‌صورت مستقیم برای بهینه‌سازی چنین معیارهای رفتاری طراحی نشده‌اند.

مسئله‌ی اصلی مورد بررسی در این حوزه آن است که چگونه می‌توان مدل‌های زبانی را به‌گونه‌ای آموزش داد که خروجی‌های آن‌ها مستقیماً با اهداف واقعی سیستم، مانند افزایش CTR، هم‌راستا شود. مقاله‌ی «Improving Reinforcement using Facebook on Text Ad Generative Learning» این مسئله را با استفاده از یادگیری تقویتی و بازخورد عملکرد واقعی کاربران بررسی می‌کند.

۲ مرور ادبیات پژوهشی

۱.۲ یادگیری تقویتی از بازخورد انسانی

یکی از مهم‌ترین کارهای پایه‌ای در هم‌راستاسازی مدل‌های زبانی، مقاله‌ی Ouyang و همکاران (۲۰۲۲) است که چارچوب Reinforcement Learning from Human Feedback (RLHF) را معرفی می‌کند. در این روش، ابتدا یک مدل پاداش با استفاده از ترجیحات انسانی آموزش داده می‌شود و سپس مدل زبانی با استفاده از الگوریتم PPO بهینه می‌گردد. این رویکرد کیفیت زبانی بالایی ایجاد می‌کند، اما وابستگی آن به بازخورد انسانی، هزینه‌بر و مقیاس‌ناپذیر است.

۲.۲ یادگیری تقویتی برای بهینه‌سازی متن

در ادامه‌ی این مسیر، Deng و همکاران (۲۰۲۲) در مقاله‌ی RLPrompt نشان دادند که یادگیری تقویتی می‌تواند برای بهینه‌سازی متون گسسته مانند هاprompt استفاده شود. اگرچه این کار اثربخشی RL در مسائل متنی را نشان می‌دهد، اما ارزیابی آن محدود به محیط‌های آزمایشگاهی است و از داده‌های واقعی کاربران استفاده نمی‌کند.

۳.۲ تحلیل انتقادی RL در NLP

Ramamurthy و همکاران (۲۰۲۲) به چالش‌هایی مانند طراحی تابع پاداش، ناپایداری آموزش و دشواری ارزیابی در استفاده از RL برای NLP اشاره می‌کنند و چارچوبی مفهومی برای تحلیل این روش‌ها ارائه می‌دهند.

۴.۲ بازخورد عملکرد واقعی

مقاله‌ی Jiang و همکاران (۲۰۲۵) روش Reinforcement Learning with Performance Feedback (RLPF) را معرفی می‌کند که در آن از داده‌های واقعی کاربران و نرخ کلیک به‌عنوان سیگنال پاداش استفاده می‌شود. این روش در یک آزمایش واقعی به بهبود معنادار عملکرد منجر شده است.

۳ جمع‌بندی

مرور ادبیات نشان می‌دهد که تحقیقات حوزه‌ی هم‌راستاسازی مدل‌های زبانی از بازخورد انسانی به سمت استفاده از سیگنال‌های مبتنی بر عملکرد واقعی حرکت کرده است. مقاله‌ی مورد بررسی نشان می‌دهد که یادگیری تقویتی می‌تواند مدل‌های زبانی را به‌طور مؤثر با اهداف واقعی سیستم‌های صنعتی هم‌راستا کند.

فهرست مقالات مرورشده

- *with Instructions Follow to Models Language Training* .(۲۰۲۲) al. et L. Ouyang، *.Feedback Human*
- *Rein- with Prompts Text Discrete Optimizing RLPrompt:* .(۲۰۲۲) al. et M. Deng، *.Learning forcement*
- *Lan- Natural for (Not) Learning Reinforcement Is* .(۲۰۲۲) al. et R. Ramamurthy، *.Processing? guage*
- *Re- using Facebook on Text Ad Generative Improving* .(۲۰۲۵) al. et R. D. Jiang، *.Learning inforcement*