



دانشگاه صنعتی شریف

مرکز تحقیقات مهندسی کوانتوم و فناوری‌های فوتونیک
گروه الگوریتم و نرم‌افزارهای کوانتومی

گزارش پروژه:

آقای احمد شفيعی

۲۷ آبان ۱۴۰۴

فهرست مطالب

۱	مفاهیم و مبانی نظری پردازش زبان طبیعی	۱
۱-۱	مقدمه	۱
۱-۱-۱	ظهور یادگیری عمیق و شبکه‌های عصبی پیچیده	۵
۱-۱-۲	مدل‌های زبانی بزرگ و تحولات اخیر	۸
۲-۱	رویکرد کوانتومی در پردازش زبان طبیعی	۱۱
۱-۲-۱	مبانی محاسبات کوانتومی	۱۱
۲-۲-۱	یادگیری ماشین کوانتومی	۱۲
۳-۲-۱	مزایای رویکرد کوانتومی در پردازش زبان طبیعی	۱۴
۴-۲-۱	چالش‌های رویکرد کوانتومی	۱۵
۵-۲-۱	وضعیت فعلی و چشم‌انداز آینده	۱۶
۳-۱	رویکردهای مختلف در پردازش زبان طبیعی کوانتومی	۱۸
۱-۳-۱	LSTM کوانتومی و شبکه‌های عصبی بازگشتی کوانتومی	۱۸
۲-۳-۱	مکانیزم‌های توجه کوانتومی	۱۹
۳-۳-۱	مدل‌های نمایش کوانتومی برای کلمات	۱۹
۴-۱	رویکرد DisCoCat: معناشناسی ترکیبی-توزیعی-مقوله‌ای	۲۰
۱-۴-۱	مبانی نظری و فلسفه رویکرد	۲۰

۲۱ دستورپیش‌گروهی	۲-۴-۱
۲۲ نمایش دیاگرامی و دیاگرام‌های رشته‌ای	۳-۴-۱
۲۳ نگاشت به فضای کوانتومی و تبدیل به مدار	۴-۴-۱
۲۴ کتابخانه لامبک : ابزار کامل برای پردازش زبان طبیعی کوانتومی	۵-۴-۱
۲۵ تحقیقات مرتبط و کاربردها	۶-۴-۱
۲۶ پردازش زبان فارسی و QNLP	۵-۱
۲۶ اهمیت و انگیزه تحقیق	۱-۵-۱
۲۷ ویژگی‌های ساختاری و نحوی زبان فارسی	۲-۵-۱
۲۸ چالش‌های پردازش زبان فارسی و QNLP	۳-۵-۱
۳۰ ابزارها و منابع موجود برای پردازش زبان فارسی	۴-۵-۱
۳۱ اهداف و چشم‌انداز این تحقیق	۵-۵-۱
۳۲ جمع‌بندی	۶-۱

فصل ۱

مفاهیم و مبانی نظری پردازش زبان طبیعی

۱-۱ مقدمه

پردازش زبان طبیعی^۲ یا NLP یکی از شاخه‌های بنیادین هوش مصنوعی است که به بررسی و پیاده‌سازی تعامل میان کامپیوتر و زبان انسانی می‌پردازد. هدف اصلی این حوزه، توانمندسازی ماشین‌ها برای درک، تفسیر و تولید زبان طبیعی به گونه‌ای است که بتوانند با انسان‌ها به صورت معنادار ارتباط برقرار کنند. این حوزه در دهه‌های اخیر شاهد تحولات شگرفی بوده است، تحولاتی که از روش‌های آماری ساده و مبتنی بر قاعده^۳ آغاز شده و به تدریج به سمت مدل‌های یادگیری عمیق پیچیده و در نهایت مدل‌های زبانی بزرگ (LLMs)^۴ گسترش یافته است.

در سال‌های اخیر، با پیشرفت‌های چشمگیر در فناوری محاسبات کوانتومی، رویکردهای نوینی برای حل مسائل پردازش زبان طبیعی ظهور کرده‌اند. این رویکردها با بهره‌برداری از ویژگی‌های منحصر به فرد مکانیک کوانتومی از جمله برهم‌نهی، درهم‌تنیدگی و تداخل^۵، امکان مدل‌سازی متفاوت و بالقوه کارآمدتری از ساختار و معنای زبان را فراهم می‌آورند. پردازش زبان طبیعی کوانتومی (QNLP)^۶ به عنوان یک پارادایم نوین، تلاش می‌کند تا با استفاده از اصول مکانیک کوانتومی، محدودیت‌های روش‌های

^۲ Natural Language Processing (NLP)

^۳ Rule-driven

^۴ Large Language Models (LLMs)

^۵ superposition, entanglement and interference

^۶ Quantum Natural Language Processing (QNLP)

کلاسیک را برطرف کرده و راه‌حلهایی با کارایی بهتر ارائه دهد.

در این فصل، ابتدا به مرور تاریخی تحولات پردازش زبان طبیعی از آغاز تا به امروز خواهیم پرداخت. سپس مدل‌های زبانی بزرگ، تأثیرات و چالش‌های آن‌ها را مورد بررسی قرار می‌دهیم. در ادامه، به معرفی رویکرد کوانتومی در پردازش زبان طبیعی، مزایا و چالش‌های آن پرداخته و رویکردهای مختلف QNLP را مقایسه می‌کنیم. به ویژه بر رویکرد ^۱ DisCoCat که مبتنی بر نظریه مقوله‌ها و دستور پیش‌گروهي است تمرکز خواهیم کرد. در نهایت، چالش‌های خاص پردازش زبان فارسی و اهداف این تحقیق را تشریح خواهیم نمود.

مقدمه‌ای بر پردازش زبان طبیعی و مسیر گذار به پردازش کوانتومی

پردازش زبان طبیعی ^۲ یا NLP یکی از شاخه‌های دیرپای هوش مصنوعی است که هدف آن ایجاد امکان درک، تولید و تحلیل زبان انسان توسط ماشین‌هاست. در دهه‌های نخست، روش‌های مبتنی بر قوانین دست‌نویس، منطق صوری و مدل‌های آماری ساده مانند مدل‌های n-gram بر این حوزه تسلط داشتند. این روش‌ها اگرچه گامی مهم در فهم ساختار زبان بودند، اما توانایی محدودی در grasp کردن پیچیدگی‌های نحوی و معنایی واقعی زبان طبیعی داشتند.

با ظهور شبکه‌های عصبی عمیق و به‌ویژه مدل‌های مبتنی بر توالی مانند RNN و LSTM، جهش قابل‌توجهی در عملکرد سامانه‌های زبانی رخ داد. این تحول با معرفی معماری Transformer و مکانیسم توجه (Attention) وارد مرحله‌ای کاملاً جدید شد که در آن مدل‌های زبانی بزرگ (LLMs) توانستند معنا، انسجام و استدلال را در ابعاد بی‌سابقه‌ای یاد بگیرند. با وجود این پیشرفت‌ها، هزینه محاسباتی بسیار بالا، نیاز عظیم به داده و محدودیت‌های بنیادی محاسبات کلاسیک، پژوهشگران را به سوی رویکردهای کارا تر سوق داده است.

در سال‌های اخیر، ایده بهره‌گیری از محاسبات کوانتومی برای پردازش زبان طبیعی - که از آن با عنوان Quantum NLP (QNLP) یاد می‌شود - مطرح شده است. این حوزه در تلاش است تا از برتری‌های کوانتومی مانند درهم‌نهی، درهم‌تنیدگی و فضا‌های هیلبرتی با ابعاد بسیار بالا برای نمایش ساختارهای زبانی بهره گیرد. استفاده از مدارهای کوانتومی، نگاهی تانسوری و مدل‌های ترکیب‌پذیر

¹ Distributional Compositional Categorical

² Natural Language Processing

معنایی می‌تواند امکان پردازش کاراتر و طبیعی‌تر زبان را در آینده فراهم کند؛ به‌ویژه در کاربردهایی مانند دسته‌بندی معنایی، ترجمه، پرسش و پاسخ و مدل‌سازی ساختار جمله.

روش DisCoCat

روش DisCoCat^۱ یکی از رویکردهای پیشرو برای مدل‌سازی زبان در چارچوب نظریه رده‌ها^۲ است. در این روش، ترکیب نحوی جمله با استفاده از گرامر پیش گروه^۳ توصیف می‌شود و این ساختار نحوی سپس به یک فضا یا مدار کوانتومی نگاشت می‌گردد که در آن معنا از طریق ترکیب تانسوری و نگاشت‌های خطی به دست می‌آید. زیبایی DisCoCat در این است که ترکیب معنایی زبان را دقیقاً موازی ترکیب‌پذیری فیزیک کوانتومی مدل می‌کند؛ به این ترتیب که نقش‌ها و وابستگی‌های نحوی مستقیماً به ساختارهای خطی در فضاهای هیلبرتی ترجمه می‌شوند. این ویژگی باعث شده که DisCoCat پایه نظری بسیاری از پژوهش‌های معاصر QNLP باشد و در چارچوب‌هایی مانند lambeq به صورت عملی و قابل اجرا پیاده‌سازی شود.

حوزه QNLP حاصل تلاش مشترک سه پژوهشگر از دانشگاه آکسفورد است: Mehrnoosh Sadrzadeh، Stephen Clark و Bob Coecke. صدرزاده بر روی جبر گرامر و ساختارهای ریاضی لازم برای مدل‌سازی نمودار زبان‌های طبیعی کار می‌کرد. کلارک به طور تخصصی روی تعبیه واژگان (Word Embedding) متمرکز بود؛ روشی که واژه‌ها را به صورت عناصر یک فضای برداری نمایش می‌دهد تا ویژگی‌های معنایی آن‌ها به طور سازگار حفظ شود. در همین زمان، کوک در حال توسعه مکانیک کوانتومی رده‌ای^۴ بود؛ چارچوبی که فرایندهای کوانتومی را به صورت ساختارهایی ترکیب‌پذیر و قابل اتصال توصیف می‌کند. ترکیب این سه خط پژوهشی زیربنای شکل‌گیری رویکرد نوین QNLP و مدل ترکیبی DisCoCat شد.

¹ Distributive Compositional Categorical Model

² Category Theory

³ Pregroup Grammar

⁴ Categorical Quantum Mechanics

مفاهیم بنیادی در چارچوب DisCoCat

گرامر (Grammar). گرامر مجموعه‌ای از قواعد صوری است که نحوه کنار هم قرار گرفتن واژه‌ها را برای تشکیل جمله مشخص می‌کند. در مدل‌های ترکیبی، گرامر نقش اسکلت ساختاری جمله را ایفا می‌کند و مبنای استخراج روابط نحوی و معنایی است.

پارسینگ (Parsing). پارسینگ فرآیند تحلیل ساختار جمله بر اساس گرامر است؛ یعنی تشخیص نقش هر کلمه و نحوه ترکیب آن‌ها با یکدیگر. برای نمونه در جمله ساده:

The cat sleeps.

فعل (sleep) به فاعل (cat) اعمال می‌شود. در جمله پیچیده‌تر:

Alice eats an apple.

پارسینگ تشخیص می‌دهد که Alice فاعل، apple مفعول و eats هسته فعلی است و این سه باید طبق الگوی نحوی استاندارد به هم متصل شوند.

پری‌گروپ (Pregroup) در برابر گروه (Group). گروه یک ساختار جبری با عمل دوتایی و وارون‌پذیری کامل است؛ یعنی هر عنصر یک معکوس دارد. اما Pregroup ساختاری ضعیف‌تر است که در آن هر نوع زبانی تنها دو شبه‌معکوس (چپ و راست) دارد و ترکیب آن‌ها با حذف‌پذیری کنترل‌شده صورت می‌گیرد. در گرامر نوعی، این ویژگی اجازه می‌دهد که وابستگی‌های نحوی مانند «فاعل-فعل» یا «فعل-مفعول» با عملیات ساده حذف نشان داده شوند.

جبر (Algebra). جبر در این زمینه به معنای مجموعه‌ای از ساختارها و قوانین ریاضی برای مدل‌کردن نحوه ترکیب اجزای زبان است. چه در سطح نحوی (مانند Pregroup) و چه در سطح معنایی (مانند ضرب تانسوری)، جبر نقش ستون فقرات محاسبات زبان را دارد.

معنای توزیعی (Distributional Semantics). این ایده بر این اصل تکیه دارد که «معنای یک واژه را می‌توان از بافت اطراف آن استنباط کرد». واژه‌ها به‌صورت بردارهایی در یک فضای برداری مدل می‌شوند و نزدیکی معنایی از طریق فاصله این بردارها سنجیده می‌شود. برای مثال بردارهای king و queen در مدل‌های توزیعی نزدیک‌اند، زیرا در محیط‌های مشابهی به کار می‌روند.

مدل‌های معناشناسی توزیعی ترکیبی.^۱ این مدل‌ها تلاش می‌کنند دو اصل را با هم ترکیب کنند: (۱) معناشناسی توزیعی برای نمایش برداری واژه‌ها، (۲) ترکیب‌پذیری نحوی برای ساخت معنای کل جمله. آن‌ها می‌کوشند معنای جمله را از طریق ترکیب مناسب بردارهای واژه‌ها و ساختار نحوی حاصل کنند.

چارچوب ترکیبی مقوله‌ای.^۲ این چارچوب از نظریه رده‌ها استفاده می‌کند تا نحوه ترکیب ساختاری (نحو) و ترکیب معنایی را در یک معماری واحد و سازگار توصیف کند. ساختارهای نحوی، نگاشت‌های خطی و ضرب تانسوری همگی به صورت فرایندهای ترکیب‌پذیر مدل می‌شوند؛ درست مانند ترکیب سیستم‌ها در مکانیک کوانتومی.

معناشناسی توزیعی چندوجهی.^۳ این رویکرد معنای واژه‌ها و جملات را تنها از متن استخراج نمی‌کند؛ بلکه از داده‌های چندگانه مانند تصویر، صدا یا تعاملات کاربری نیز بهره می‌گیرد. در این روش، فضاهای برداری چندوجهی ساخته می‌شوند که معنا را از چند منبع هم‌زمان یاد می‌گیرند؛ برای مثال ترکیب تصویر یک «گربه» با توصیف متنی آن برای بهبود معنای برداری واژه cat.

۱-۱-۱ ظهور یادگیری عمیق و شبکه‌های عصبی پیچیده

با افزایش قدرت محاسباتی سخت‌افزارها، به ویژه واحدهای پردازش گرافیکی (GPUs)^۴، و در دسترس بودن مجموعه داده‌های بزرگ، استفاده از شبکه‌های عصبی عمیق از اواسط دهه ۲۰۱۰ به طور گسترده در پردازش زبان طبیعی رواج یافت. این دوره شاهد پیشرفت‌های چشمگیری در معماری‌های شبکه‌های عصبی بود که توانایی مدل‌سازی پیچیدگی‌های زبان را به طور قابل ملاحظه‌ای افزایش دادند.

یکی از اولین معماری‌های موفق در این دوره، شبکه‌های عصبی بازگشتی (RNN)^۵ بود. این شبکه‌ها با داشتن حلقه‌های بازخورد، قادر بودند اطلاعات از گام‌های قبلی را حفظ کنند و از این رو برای پردازش داده‌های توالی‌ای مانند متن مناسب بودند. با این حال، RNN‌های ساده با مشکل محو شدن گرادیان یا^۶ مواجه بودند که مانع از یادگیری وابستگی‌های بلندمدت می‌شد.

¹ Compositional Distributional Semantics Models

² Categorical Compositional Framework

³ Multimodal Distributional Semantics

⁴ Graphics Processing Units

⁵ Recurrent Neural Networks

⁶ Vanishing gradient

برای حل این مشکل، هوخرایتر و اشمیدهور در سال ۱۹۹۷ معماری حافظه طولانی کوتاه‌مدت- (LSTM) [۳]^۱ را معرفی کردند. اگرچه این معماری در اواخر دهه ۱۹۹۰ ابداع شد، اما استفاده گسترده از آن در پردازش زبان طبیعی در دهه ۲۰۱۰ آغاز شد. LSTM با استفاده از ساختار پیچیده‌تری شامل دروازه‌های مختلف، توانست اطلاعات بلندمدت را به طور مؤثر حفظ کند. این معماری دارای سه دروازه اصلی است: دروازه فراموشی که تصمیم می‌گیرد چه اطلاعاتی از حافظه حذف شود، دروازه ورودی که کنترل می‌کند چه اطلاعات جدیدی به حافظه اضافه شود، و دروازه خروجی^۲ که تعیین می‌کند چه بخشی از حافظه به عنوان خروجی ارائه شود. علاوه بر این، یک حالت سلول^۳ وجود دارد که به عنوان حافظه بلندمدت عمل می‌کند.

شبکه‌های LSTM و نسخه ساده‌تر آن به نام واحدهای بازگشتی دارای دروازه-GRU^۴ در بسیاری از وظایف پردازش زبان طبیعی از جمله ترجمه ماشینی، تحلیل احساسات و تولید متن به کار گرفته شدند و نتایج قابل توجهی به دست آوردند.

تحول عمده بعدی در سال ۲۰۱۴ با معرفی مکانیزم توجه^۵ توسط بهدانا و همکاران رخ داد. این مکانیزم به مدل اجازه می‌داد تا به جای اینکه تنها بر یک بردار زمینه ثابت تکیه کند، به بخش‌های مختلف ورودی با وزن‌های متفاوت توجه کند. این نوآوری به ویژه در ترجمه ماشینی عصبی تأثیر بسزایی داشت. اوج این تحولات با انتشار مقاله تأثیرگذار "توجه تنها چیزی است که نیاز دارید"^۶ [۴] توسط واسوانی و همکاران در سال ۲۰۱۷ رخ داد. این مقاله معماری ترنسفورمر^۷ را معرفی کرد که به طور کامل بر مکانیزم توجه استوار است و نیازی به شبکه‌های بازگشتی ندارد. مکانیزم توجه در ترنسفورمر به صورت زیر تعریف می‌شود:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (1-1)$$

¹ Long Short-Term Memory

² Forget Gate, Input Gate and Output Gate

³ Cell State

⁴ Gated Recurrent Units

⁵ Attention Mechanism

⁶ Attention Is All You Need

⁷ Transformer

در این رابطه، Q نمایانگر پرس وجوها، K نمایانگر کلیدها و V نمایانگر مقادیر^۱ است. ضرب $\sqrt{d_k}$ برای مقیاس بندی استفاده می شود تا از مقادیر بسیار بزرگ در تابع softmax جلوگیری شود.

معماری ترنسفورمر مزایای قابل توجهی نسبت به معماری های قبلی داشت. اول اینکه، امکان پردازش موازی تمام توکن های ورودی را فراهم می کرد، در حالی که RNN ها به دلیل ماهیت توالی ای شان باید به صورت ترتیبی پردازش شوند. این ویژگی آموزش را بسیار سریع تر می کرد. دوم، مکانیزم توجه چندسره^۲ به مدل اجازه می داد تا همزمان به جنبه های مختلف رابطه میان کلمات توجه کند. سوم اینکه، این معماری توانایی بهتری در مدل سازی وابستگی های بلندمدت داشت و مقیاس پذیری بسیار بالایی را ارائه می کرد. معماری ترنسفورمر به سرعت به یک استاندارد در پردازش زبان طبیعی تبدیل شد و پایه و اساس تمام مدل های زبانی بزرگ بعدی را تشکیل داد.

این مقاله که به یکی از پراستنادترین آثار علمی در تاریخ هوش مصنوعی تبدیل شده است، راه حلی جسورانه برای محدودیت های شبکه های بازگشتی ارائه داد. قبل از معرفی ترنسفورمر، RNN ها و LSTM ها علی رغم موفقیت هایشان با دو مشکل اساسی مواجه بودند: عدم امکان موازی سازی محاسبات به دلیل ماهیت ترتیبی پردازش، و ضعف در یادگیری وابستگی های بسیار بلندمدت در جملات طولانی. واسوانی و همکارانش با حذف کامل ساختارهای بازگشتی و تکیه صرف بر مکانیزم توجه، این محدودیت ها را برطرف کردند. معماری ترنسفورمر بر پایه دو بخش اصلی کدگذار و رمزگشا ساخته شده که هر کدام از لایه های مکانیزم توجه چندسره و شبکه های پیشخور تشکیل شده اند.

قلب این معماری، مکانیزم توجه مقیاس شده ضرب نقطه ای است که بر اساس سه مفهوم پرس وجو، کلید و مقدار عمل می کند و به مدل اجازه می دهد به طور دینامیک تصمیم بگیرد کدام بخش های ورودی برای پردازش هر کلمه اهمیت دارند. استفاده از توجه چندسره به مدل امکان می دهد همزمان به جنبه های مختلف رابطه میان کلمات از جمله وابستگی های نحوی، روابط معنایی و همپوشانی های واژگانی توجه کند. برای حفظ اطلاعات مربوط به ترتیب کلمات که با حذف ساختارهای بازگشتی از دست می رفت، کدگذاری موقعیت بر اساس توابع سینوسی و کسینوسی به مدل اضافه شد.

نتایج آزمایش های این مقاله در ترجمه ماشینی چشمگیر بود، به گونه ای که نه تنها رکوردهای قبلی را شکست بلکه زمان آموزش را از هفته ها به تنها چند روز کاهش داد. اما تأثیر واقعی این مقاله فراتر از نتایج عددی بود. ترنسفورمر به سرعت پایه تمام مدل های بزرگ بعدی شد: BERT که تنها از بخش

¹ Queries, Keys and Values

² Multi-Head Attention

کدگذار استفاده کرد و با پیش‌آموزش دوطرفه رکوردهای متعددی شکست، سری مدل‌های GPT که از بخش رمزگشا بهره بردند و در تولید متن طبیعی توانایی شگفت‌انگیز نشان دادند، و در نهایت GPT-3 با ۱۷۵ میلیارد پارامتر که نقطه عطفی در تاریخ هوش مصنوعی شد. امروزه تقریباً تمام سیستم‌های پیشرفته پردازش زبان از مترجم‌های خودکار تا دستیارهای هوشمند مانند ChatGPT بر پایه این معماری استوارند، و حتی فراتر از پردازش زبان، در بینایی کامپیوتر، زیست‌شناسی محاسباتی و پیش‌بینی ساختار پروتئین‌ها نیز کاربرد یافته است. در [پیوست؟؟]، تحلیل جامع‌تری از این مقاله تأثیرگذار و نقش آن در تحول هوش مصنوعی زبانی ارائه شده است.

۱-۱-۲ مدل‌های زبانی بزرگ و تحولات اخیر

پس از معرفی معماری ترنسفورمر، تحقیقات به سمت ساخت مدل‌های بزرگ‌تر و قدرتمندتر پیش رفت. ایده اصلی این بود که با افزایش اندازه مدل و حجم داده‌های آموزشی، عملکرد مدل در طیف وسیعی از وظایف بهبود می‌یابد. این رویکرد به ظهور مدل‌های زبانی بزرگ-LLMs^۱ منجر شد که توانایی‌های شگفت‌انگیزی در فهم و تولید زبان طبیعی از خود نشان دادند.

یکی از اولین و تأثیرگذارترین مدل‌های این دوره، BERT^۲ [۵] بود که در اکتبر ۲۰۱۸ توسط محققان گوگل معرفی شد. نوآوری اصلی BERT در استفاده از پیش‌آموزش دوطرفه بود، به این معنا که مدل هنگام پیش‌بینی یک کلمه، هم به کلمات قبل و هم به کلمات بعد از آن توجه می‌کند. این رویکرد برخلاف مدل‌های زبانی سنتی بود که تنها به صورت یک‌طرفه (از چپ به راست یا بالعکس) عمل می‌کردند.

BERT از دو وظیفه پیش‌آموزشی استفاده می‌کند: اول، مدل‌سازی زبانی پوشیده^۳ (MLM) که در آن برخی کلمات به صورت تصادفی پوشیده شده و مدل باید آن‌ها را پیش‌بینی کند. دوم، پیش‌بینی جمله بعدی-NSP^۴ که مدل باید تشخیص دهد آیا دو جمله در متن اصلی پشت سر هم آمده‌اند یا خیر. پس از پیش‌آموزش بر روی پیکره‌های عظیم متنی، BERT می‌تواند با تنظیم ظریف^۵ بر روی مجموعه داده‌های کوچک‌تر برای وظایف خاص مانند پاسخ به پرسش، تحلیل احساسات و شناسایی موجودیت‌های اسمی

^۱ Large Language Models

^۲ Bidirectional Encoder Representations from Transformers

^۳ Masked Language Modeling

^۴ Next Sentence Prediction

^۵ Fine-tuning

آموزش ببیند. این رویکرد منجر به شکستن رکوردهای متعددی در معیارهای سنجش استاندارد^۱ پردازش زبان طبیعی شد.

در مسیر موازی، مدل‌های تولیدی^۲ نیز در حال پیشرفت بودند. سری مدل‌های GPT^۳ توسط شرکت OpenAI توسعه یافت. GPT-1 در سال ۲۰۱۸ معرفی شد و نشان داد که یک مدل زبانی بزرگ که بر روی پیکره گسترده‌ای پیش‌آموزش دیده، می‌تواند با حداقل تنظیم در وظایف مختلف عملکرد خوبی داشته باشد.

GPT-2 در سال ۲۰۱۹ با ۵.۱ میلیارد پارامتر معرفی شد و توانایی تولید متن‌های منسجم و طبیعی را نشان داد. در سال ۲۰۲۰، GPT-3 [۶] با ۱۷۵ میلیارد پارامتر، نقطه عطفی در تاریخ پردازش زبان طبیعی بود. این مدل توانایی یادگیری چندوظیفه‌ای^۴ را به نمایش گذاشت، به این معنا که با دیدن تنها چند مثال از یک وظیفه جدید، می‌توانست آن وظیفه را انجام دهد بدون اینکه نیازی به تنظیم ظریف پارامترهایش باشد. حتی در حالت یادگیری بدون مثال^۵، GPT-3 در بسیاری از وظایف عملکرد قابل قبولی داشت.

در سال‌های اخیر، مدل‌های پیشرفته‌تری نیز معرفی شده‌اند. GPT-4 که در سال ۲۰۲۳ عرضه شد، توانایی‌های چندوجهی^۶ دارد و می‌تواند علاوه بر متن، با تصاویر نیز کار کند. مدل‌های دیگری مانند PaLM از گوگل، LLaMA از متا، و Claude از Anthropic نیز در این حوزه فعالیت می‌کنند. این مدل‌ها در کاربردهای متنوعی از ترجمه ماشینی گرفته تا کدنویسی خودکار، تحلیل داده، و حتی تحقیقات علمی مورد استفاده قرار می‌گیرند.

مدل‌های زبانی بزرگ تأثیرات عمیق و گسترده‌ای بر جامعه، صنعت و تحقیقات علمی داشته‌اند. بهبود چشمگیر در کیفیت ترجمه ماشینی، یکی از بارزترین دستاوردهای این مدل‌ها است. سیستم‌های ترجمه مبتنی بر LLMs توانسته‌اند کیفیت ترجمه را به سطحی برسانند که در بسیاری موارد با ترجمه انسانی قابل مقایسه است. دستیارهای هوشمند و چت‌بات‌های پیشرفته مانند ChatGPT، Gemini و Claude توانسته‌اند تعاملات طبیعی و معناداری با کاربران برقرار کنند و در حوزه‌های مختلف از پشتیبانی مشتری تا آموزش و یادگیری کمک کنند.

¹ benchmarks

² Generative models

³ Generative Pre-trained Transformer

⁴ Few-shot learning

⁵ zero-shot learning

⁶ multimodal

خلاصه‌سازی خودکار اسناد، تولید محتوای متنی، نوشتن کد برنامه‌نویسی، و کمک به تحقیقات علمی و پزشکی از دیگر کاربردهای مهم این مدل‌ها هستند. در حوزه تحقیقات، LLMs می‌توانند در بررسی ادبیات، تحلیل داده‌ها، و حتی پیشنهاد فرضیه‌های جدید به محققان کمک کنند.

با این حال، مدل‌های زبانی بزرگ با چالش‌ها و مسائل جدی‌ای نیز مواجه هستند که نیازمند توجه و تحقیقات بیشتر است. یکی از اساسی‌ترین چالش‌ها، نیاز به منابع محاسباتی عظیم برای آموزش و اجرای این مدل‌هاست. به عنوان مثال، آموزش GPT-3 با ۱۷۵ میلیارد پارامتر نیازمند هزاران ساعت محاسبه بر روی صدها پردازنده گرافیکی قدرتمند بود که هزینه‌ای در حدود میلیون‌ها دلار دارد. این امر دسترسی به این فناوری را محدود به شرکت‌ها و سازمان‌های بزرگ می‌کند و شکاف دیجیتال را تشدید می‌کند.

مصرف انرژی و اثرات زیست‌محیطی این مدل‌ها نیز نگرانی جدی است. آموزش یک مدل زبانی بزرگ می‌تواند ردپای کربنی قابل توجهی داشته باشد که معادل سوخت مصرفی چندین خودرو در طول عمرشان است. این موضوع اهمیت توسعه الگوریتم‌های کارآمدتر و استفاده از انرژی‌های تجدیدپذیر را برجسته می‌کند.

تعصب و انصاف^۱ یکی دیگر از چالش‌های اساسی است. این مدل‌ها بر روی داده‌های گسترده‌ای از اینترنت آموزش می‌بینند که ممکن است حاوی تعصبات اجتماعی، نژادی، جنسیتی و فرهنگی باشند. در نتیجه، مدل نیز این تعصبات را یاد می‌گیرد و در خروجی‌هایش بازتاب می‌دهد. این موضوع می‌تواند منجر به تصمیم‌گیری‌های ناعادلانه در کاربردهای حساس مانند استخدام، اعتباردهی و قضاوت‌های قانونی شود.

قابلیت تفسیر^۲ محدود این مدل‌ها نیز چالش مهمی است. مدل‌های زبانی بزرگ به صورت "جعبه سیاه" عمل می‌کنند، به این معنا که حتی طراحان آن‌ها نمی‌توانند به طور دقیق توضیح دهند که چرا مدل یک پاسخ خاص را تولید کرده است. این عدم شفافیت در کاربردهای بحرانی مانند تشخیص پزشکی یا تصمیمات قانونی می‌تواند مشکل‌ساز باشد.

یکی از مشکلات شناخته شده LLMs، پدیده توهم یا هذیان‌گویی^۳ است. این مدل‌ها گاهی اطلاعات نادرست یا جعلی را با اطمینان و قاطعیت بیان می‌کنند، به گونه‌ای که تشخیص درستی یا نادرستی آن‌ها برای کاربر عادی دشوار است. این موضوع می‌تواند منجر به انتشار اطلاعات غلط و

^۱ bias and fairness

^۲ Interpretability

^۳ Hallucination

گمراه‌کنندگی شود.

علاوه بر این، مسائل امنیتی، حریم خصوصی، حقوق مالکیت فکری، و سوءاستفاده احتمالی از این فناوری برای تولید محتوای مخرب یا گمراه‌کننده نیز از دیگر چالش‌های مهم هستند که نیازمند راهکارهای فنی، قانونی و اخلاقی هستند.

۲-۱ رویکرد کوانتومی در پردازش زبان طبیعی

با وجود پیشرفت‌های چشمگیر در مدل‌های زبانی بزرگ کلاسیک، محدودیت‌ها و چالش‌های ذکر شده در بخش قبل انگیزه‌ای برای جستجوی رویکردهای جایگزین یا مکمل شده است. یکی از رویکردهای نوظهور که در سال‌های اخیر توجه محققان را به خود جلب کرده، استفاده از محاسبات کوانتومی در پردازش زبان طبیعی است. این رویکرد با بهره‌گیری از اصول بنیادین مکانیک کوانتومی، چشم‌اندازهای جدیدی را برای مدل‌سازی زبان و معنا فراهم می‌آورد.

۱-۲-۱ مبانی محاسبات کوانتومی

برای درک رویکرد کوانتومی در پردازش زبان طبیعی، ابتدا لازم است با مفاهیم پایه‌ای محاسبات کوانتومی آشنا شویم. محاسبات کوانتومی بر اساس اصول مکانیک کوانتومی عمل می‌کند و از ویژگی‌هایی بهره می‌برد که در محاسبات کلاسیک وجود ندارند.

واحد اطلاعات کلاسیک، بیت است که می‌تواند دو مقدار ۰ یا ۱ داشته باشد. در محاسبات کوانتومی، واحد اساسی اطلاعات، کیوبیت نامیده می‌شود. یک کیوبیت می‌تواند در حالت برهم‌نهی قرار گیرد، به این معنا که به صورت همزمان در ترکیبی از هر دو حالت $|0\rangle$ و $|1\rangle$ باشد. این خاصیت به صورت زیر نمایش داده می‌شود:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \quad (۲-۱)$$

در این رابطه، α و β اعداد مختلط هستند که در آن‌ها $|\alpha|^2$ احتمال مشاهده حالت $|0\rangle$ و $|\beta|^2$ احتمال مشاهده حالت $|1\rangle$ را نشان می‌دهد. قید نرمال‌سازی $|\alpha|^2 + |\beta|^2 = 1$ باید برقرار باشد. این ویژگی به

یک کیوبیت اجازه می‌دهد تا اطلاعات بیشتری نسبت به یک بیت کلاسیک نمایش دهد.

یکی دیگر از ویژگی‌های بنیادین مکانیک کوانتومی، درهم‌تنیدگی است. وقتی دو یا چند کیوبیت درهم‌تنیده می‌شوند، حالت یک کیوبیت به طور جدایی‌ناپذیری به حالت کیوبیت‌های دیگر وابسته می‌شود، به گونه‌ای که نمی‌توان آن‌ها را به صورت مستقل توصیف کرد. یک مثال کلاسیک از حالت درهم‌تنیده، حالت بل است:

$$|\Phi^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle) \quad (3-1)$$

در این حالت، اگر کیوبیت اول را اندازه‌گیری کنیم و نتیجه ۰ باشد، کیوبیت دوم نیز قطعاً ۰ خواهد بود، و اگر کیوبیت اول ۱ باشد، کیوبیت دوم نیز ۱ خواهد بود. این همبستگی قوی، بدون اینکه هیچ اطلاعاتی میان کیوبیت‌ها منتقل شود، برقرار است و یکی از عجیب‌ترین پدیده‌های مکانیک کوانتومی محسوب می‌شود.

ویژگی سوم که در محاسبات کوانتومی نقش مهمی دارد، تداخل^۱ است. در مکانیک کوانتومی، دامنه‌های احتمال می‌توانند با یکدیگر تداخل کنند. تداخل سازنده باعث افزایش احتمال یک پیامد خاص می‌شود، در حالی که تداخل مخرب احتمال را کاهش می‌دهد. الگوریتم‌های کوانتومی از این ویژگی برای تقویت احتمال پاسخ‌های صحیح و کاهش احتمال پاسخ‌های نادرست استفاده می‌کنند.

این سه ویژگی - درهم‌تنیدگی و تداخل - پایه و اساس قدرت محاسباتی کامپیوترهای کوانتومی را تشکیل می‌دهند و امکانات جدیدی برای حل مسائل پیچیده فراهم می‌آورند.

۱-۲-۲ یادگیری ماشین کوانتومی

یادگیری ماشین کوانتومی (QML)^۲ [۷، ۸] به کاربرد الگوریتم‌ها و مدارات کوانتومی در حل مسائل یادگیری ماشین می‌پردازد. این حوزه در تلاقی دو رشته یادگیری ماشین و محاسبات کوانتومی قرار دارد و هدف آن بهره‌گیری از مزایای هر دو است.

در عصر فعلی که (NISQ)^۳ نامیده می‌شود، کامپیوترهای کوانتومی با تعداد محدود کیوبیت

¹ Interference

² Quantum Machine Learning

³ Noisy Intermediate-Scale Quantum

(معمولاً کمتر از چند صد) و با نویز قابل توجه در دسترس هستند. برای استفاده مؤثر از این سخت‌افزارهای محدود، الگوریتم‌های وردشی کوانتومی^۱ (VQA) توسعه یافته‌اند. این الگوریتم‌ها ترکیبی از پردازش کوانتومی و کلاسیک هستند و از مدارات کوانتومی پارامتری (PQC)^۲ استفاده می‌کنند.

ساختار کلی یک الگوریتم تغییراتی کوانتومی به صورت زیر است: ابتدا داده‌های ورودی کلاسیک به یک حالت کوانتومی تبدیل می‌شوند، این فرآیند را کدگذاری^۳ می‌نامند. سپس یک مدار کوانتومی پارامتری بر روی این حالت اعمال می‌شود که شامل دروازه‌های کوانتومی با پارامترهای قابل تنظیم است. این پارامترها معمولاً زوایای چرخش دروازه‌های کوانتومی هستند. در نهایت، حالت کوانتومی نهایی اندازه‌گیری شده و به خروجی کلاسیک تبدیل می‌شود. یک بهینه‌ساز کلاسیک، پارامترهای مدار را به منظور کمینه کردن یک تابع هزینه تنظیم می‌کند.

نگاشت ویژگی‌های کوانتومی^۴ نقش مهمی در یادگیری ماشین کوانتومی دارد. این نگاشت‌ها داده‌های کلاسیک را به فضای هیلبرت کوانتومی با ابعاد بسیار بالا منتقل می‌کنند. در این فضای با ابعاد بالاتر، الگوهایی که در فضای کلاسیک خطی قابل تفکیک نیستند ممکن است قابل تفکیک شوند. این ایده مشابه استفاده از هسته یا kernel در ماشین‌های بردار پشتیبان (SVM)^۵ است.

مفهوم Ansatz در محاسبات کوانتومی اهمیت ویژه‌ای دارد. Ansatz به معنای یک حدس آموزش‌یافته یا ساختار اولیه است که برای حل یک مسئله انتخاب می‌شود. در زمینه یادگیری ماشین کوانتومی، Ansatz به ساختار و آرایش دروازه‌های کوانتومی در مدار پارامتری اشاره دارد. انتخاب Ansatz مناسب تأثیر مستقیم بر عملکرد مدل دارد. برخی از Ansatz های معروف عبارتند از: Hardware Efficient Ansatz که با توجه به محدودیت‌های سخت‌افزار طراحی می‌شود، IQP Ansatz^۶ که در کاربردهای QNLP کاربرد دارد، و UCCSD Ansatz که در شیمی کوانتومی استفاده می‌شود.

در زمینه پردازش زبان طبیعی، یادگیری ماشین کوانتومی می‌تواند به روش‌های مختلفی به کار گرفته شود. یکی از رویکردها، استفاده از مدارات کوانتومی برای یادگیری نمایش‌های برداری کلمات است. روش دیگر، ساخت طبقه‌بندهای کوانتومی برای وظایف مانند تحلیل احساسات یا دسته‌بندی متن است که در آن معنای کلی یک جمله به یک حالت کوانتومی نگاشت می‌شود و سپس اندازه‌گیری برای تعیین

¹ Variational Quantum Algorithms

² Parameterized Quantum Circuits

³ Encoding

⁴ Quantum Feature Maps

⁵ Support Vector Machines

⁶ Instantaneous Quantum Polynomial Ansatz

برچسب انجام می‌گیرد.

۱-۲-۳ مزایای رویکرد کوانتومی در پردازش زبان طبیعی

رویکرد کوانتومی در پردازش زبان طبیعی مزایای ذاتی و منحصر به فردی دارد که از ویژگی‌های بنیادین مکانیک کوانتومی ناشی می‌شود.

یکی از مهم‌ترین مزایا، نمایش طبیعی ترکیب‌پذیری^۱ زبان است. زبان طبیعی ذاتاً ترکیبی است، به این معنا که معنای یک جمله از معنای اجزای آن و نحوه ترکیب آن‌ها حاصل می‌شود. در مکانیک کوانتومی، ترکیب سیستم‌ها از طریق ضرب تانسوری یا ضرب تانسوری^۲ در فضای هیلبرت انجام می‌شود. این تناظر طبیعی میان ساختار ترکیبی زبان و ساختار ریاضی فضای هیلبرت، امکان مدل‌سازی مستقیم و شهودی از معناشناسی زبان را فراهم می‌آورد. به عنوان مثال، معنای عبارت "سیب قرمز" می‌تواند به صورت ترکیب تانسوری حالت‌های کوانتومی مربوط به "سیب" و "قرمز" نمایش داده شود.

مزیت دیگر، توانایی مدل‌سازی ابهام و چندمعنایی است. در زبان طبیعی، بسیاری از کلمات و جملات دارای معانی متعدد هستند که بسته به زمینه تعیین می‌شوند. برهم‌نهی کوانتومی به طور طبیعی می‌تواند این ابهامات معنایی را نمایش دهد. یک کلمه چندمعنا می‌تواند به عنوان یک برهم‌نهی از حالت‌های مختلف معنایی مدل شود، و فرآیند درک زمینه می‌تواند به عنوان فروپاشی^۳ این برهم‌نهی به یک معنای خاص در نظر گرفته شود.

کاهش ابعاد مورد نیاز نیز از مزایای قابل توجه است. فضای هیلبرت کوانتومی ابعاد نمایی نسبت به تعداد کیوبیت دارد. به عنوان مثال، n کیوبیت می‌تواند فضایی با 2^n بعد را نمایش دهد. این ویژگی امکان نمایش معانی پیچیده و روابط معنایی را با تعداد نسبتاً کمی از کیوبیت فراهم می‌آورد، در حالی که در روش‌های کلاسیک ممکن است به بردارهای با ابعاد بسیار بالا نیاز باشد.

درهم‌تنیدگی کوانتومی قابلیت منحصر به فردی برای مدل‌سازی همبستگی‌های غیرکلاسیک فراهم می‌کند. در زبان طبیعی، کلمات و مفاهیم اغلب به گونه‌ای با یکدیگر مرتبط هستند که نمی‌توان آن‌ها را به صورت مستقل توصیف کرد. برای مثال، در عبارت "شاه و ملکه"، این دو مفهوم به شدت به یکدیگر

¹ compositionality

² tensor product

³ collapse

وابسته‌اند و درهم‌تنیدگی می‌تواند این وابستگی معنایی پیچیده را به طور طبیعی نمایش دهد.

علاوه بر این، رویکرد کوانتومی پتانسیل کارایی محاسباتی در برخی وظایف را دارد. اگرچه این موضوع هنوز در مراحل اولیه تحقیق است، اما امید می‌رود که در آینده الگوریتم‌های کوانتومی بتوانند برخی محاسبات مربوط به پردازش زبان را سریع‌تر از روش‌های کلاسیک انجام دهند، به ویژه در مواردی که نیاز به جستجو در فضای حالت بزرگ است.

۱-۲-۴ چالش‌های رویکرد کوانتومی

با وجود مزایای مهم رویکرد کوانتومی، این حوزه با چالش‌های جدی و بنیادینی مواجه است که بر توسعه و کاربرد عملی آن تأثیر می‌گذارد.

یکی از اساسی‌ترین چالش‌ها، محدودیت‌های سخت‌افزاری فعلی است. در حال حاضر ما در عصر کوانتومی میان‌مقیاس پرنویز یا عصر کوانتومی میان‌مقیاس پرنویز^۱ قرار داریم. کامپیوترهای کوانتومی موجود دارای تعداد محدودی کیوبیت هستند (معمولاً بین ۵۰ تا چند صد کیوبیت) و این کیوبیت‌ها دچار نویز و خطای قابل توجهی هستند. این خطاها از منابع مختلفی مانند نقص در دروازه‌های کوانتومی، خطای اندازه‌گیری، و تعامل ناخواسته با محیط ناشی می‌شوند. در نتیجه، نتایج محاسبات کوانتومی همیشه دقیق نیستند و نیاز به تکرار آزمایش‌ها و متوسط‌گیری نتایج وجود دارد.

واهلیدگی^۲ یکی دیگر از چالش‌های اساسی است. سیستم‌های کوانتومی بسیار حساس به تعاملات با محیط اطرافشان هستند. هر تعامل ناخواسته با محیط می‌تواند باعث از دست رفتن خواص کوانتومی مانند برهم‌نهی و درهم‌تنیدگی شود. این فرآیند که واهلیدگی نامیده می‌شود، معمولاً در مقیاس‌های زمانی بسیار کوتاه (میکروثانیه تا میلی‌ثانیه) رخ می‌دهد. بنابراین، محاسبات کوانتومی باید به سرعت انجام شوند و عمق مدار^۳ (تعداد لایه‌های دروازه‌های کوانتومی) باید محدود باشد. این محدودیت، پیچیدگی مدارات قابل اجرا را کاهش می‌دهد.

محدودیت تعداد کیوبیت نیز چالش عملی مهمی است. بسیاری از مسائل واقعی پردازش زبان طبیعی نیازمند پردازش جملات طولانی یا اسناد بزرگ هستند. نگاشت این مسائل به مدارات کوانتومی

^۱ NISQ Era (Noisy Intermediate-Scale Quantum)

^۲ decoherence

^۳ circuit depth

ممکن است به تعداد زیادی کیوبیت نیاز داشته باشد که فراتر از ظرفیت سخت افزارهای فعلی است. این محدودیت باعث می شود که تحقیقات فعلی عمدتاً بر روی مسائل کوچک و نمونه های اثبات مفهوم یا اثبات مفهوم^۱ متمرکز شوند.

مقیاس پذیری^۲ چالش دیگری است. حتی اگر بتوانیم یک وظیفه ساده پردازش زبان طبیعی را بر روی کامپیوتر کوانتومی انجام دهیم، گسترش آن به مسائل واقعی و پیچیده دشوار است. بسیاری از کاربردهای واقعی نیازمند پردازش میلیون ها کلمه، پیکره های بزرگ متنی، و مدل های با میلیارد ها پارامتر هستند. نحوه مقیاس دهی رویکردهای کوانتومی به این ابعاد، هنوز یک سؤال باز است.

ارتباط میان دنیای کوانتومی و کلاسیک نیز چالش برانگیز است. داده های ورودی (متن) و خروجی (مثلاً برچسب دسته بندی) معمولاً کلاسیک هستند و باید به حالت های کوانتومی تبدیل و سپس دوباره به اطلاعات کلاسیک برگردانده شوند. این فرآیند کدگذاری و رمزگشایی می تواند هزینه بر باشد. علاوه بر این، اندازه گیری کوانتومی ذاتاً احتمالی است و اطلاعات کامل حالت کوانتومی را از بین می برد، که ممکن است منجر به از دست رفتن بخشی از اطلاعات معنایی شود.

در نهایت، کمبود الگوریتم ها و روش های بهینه برای کاربردهای پردازش زبان طبیعی کوانتومی نیز یک چالش است. بسیاری از روش های موجود هنوز در مراحل اولیه تحقیق هستند و نیاز به توسعه بیشتر دارند. همچنین، مقایسه عادلانه میان روش های کوانتومی و کلاسیک دشوار است، زیرا سخت افزار کوانتومی هنوز به بلوغ نرسیده و نمی توان قضاوت قطعی در مورد برتری یا عدم برتری آن در مقایسه با روش های کلاسیک بهینه شده کرد.

۱-۲-۵ وضعیت فعلی و چشم انداز آینده

در حال حاضر، پردازش زبان طبیعی کوانتومی در مرحله تحقیقات بنیادین و آزمایش های اثبات مفهوم قرار دارد. محققان در دانشگاه ها و شرکت های فناوری در حال کاوش در پتانسیل ها و محدودیت های این رویکرد هستند. پلتفرم های ابری کوانتومی مانند IBM Quantum Experience، Google Quantum AI، Amazon Braket، و IonQ امکان دسترسی به کامپیوترهای کوانتومی واقعی را برای محققان فراهم کرده اند، که این امر باعث تسریع تحقیقات در این حوزه شده است.

¹ proof-of-concept

² scalability

اکثر کارهای تحقیقاتی فعلی بر روی وظایف ساده پردازش زبان طبیعی مانند دسته‌بندی جملات کوتاه، تحلیل احساسات پایه، و سایر مسائل با مجموعه داده‌های کوچک متمرکز است. این محدودیت عمدتاً به دلیل قید و بندهای سخت‌افزاری کنونی است. با این حال، نتایج اولیه نشان‌دهنده امکان‌پذیری رویکرد و پتانسیل‌های امیدوارکننده آن هستند.

چشم‌انداز آینده پردازش زبان طبیعی کوانتومی به عوامل متعددی بستگی دارد. یکی از مهم‌ترین آن‌ها، پیشرفت در سخت‌افزار کوانتومی است. انتظار می‌رود که در دهه‌های آینده، کامپیوترهای کوانتومی با تعداد کیوبیت بیشتر، نرخ خطای پایین‌تر، و زمان‌های انسجام طولانی‌تر در دسترس قرار گیرند. توسعه کدهای تصحیح خطای کوانتومی^۱ نیز می‌تواند به ساخت کامپیوترهای کوانتومی قابل اعتمادتر کمک کند، که این امر مسیر را برای اجرای الگوریتم‌های پیچیده‌تر هموار می‌کند.

از سوی دیگر، توسعه الگوریتم‌های الهام‌گرفته از کوانتوم^۲ برای سیستم‌های کلاسیک نیز مسیر امیدوارکننده‌ای است. این الگوریتم‌ها از ایده‌های کوانتومی مانند ضرب تانسوری و ساختارهای ترکیبی استفاده می‌کنند اما بر روی کامپیوترهای کلاسیک اجرا می‌شوند. ممکن است این رویکردها بتوانند بدون نیاز به سخت‌افزار کوانتومی، برخی مزایای رویکرد کوانتومی را به ارمغان بیاورند.

رویکرد ترکیبی کوانتومی-کلاسیک^۳ احتمالاً واقعی‌ترین مسیر برای کاربردهای نزدیک آینده است. در این رویکرد، بخش‌هایی از محاسبات که از مزایای کوانتومی بهره‌مند می‌شوند بر روی پردازنده کوانتومی اجرا می‌شوند، در حالی که بقیه محاسبات توسط رایانه‌های کلاسیک کارآمد انجام می‌شود. این تقسیم کار می‌تواند به استفاده بهینه از منابع محدود کوانتومی منجر شود.

همچنین انتظار می‌رود که کاربردهای تخصصی در حوزه‌های خاص پردازش زبان طبیعی توسعه یابند. به جای تلاش برای جایگزینی کامل مدل‌های کلاسیک، رویکرد کوانتومی ممکن است در وظایف خاصی که با ویژگی‌های کوانتومی سازگاری بیشتری دارند، مزیت رقابتی پیدا کند. برای مثال، وظایفی که نیازمند مدل‌سازی روابط پیچیده معنایی یا ساختارهای بسیار ترکیبی هستند، ممکن است کاندیدهای مناسبی برای رویکرد کوانتومی باشند.

^۱ Quantum Error Correction (QEC)

^۲ quantum-inspired algorithms

^۳ Hybrid Quantum-Classical

۳-۱ رویکردهای مختلف در پردازش زبان طبیعی کوانتومی

محققان رویکردهای متنوعی را برای ترکیب محاسبات کوانتومی و پردازش زبان طبیعی پیشنهاد کرده‌اند. این رویکردها از نظر میزان اتکا به اصول کوانتومی، نحوه مدل‌سازی ساختار زبان، و استراتژی‌های بهینه‌سازی با یکدیگر متفاوت هستند. در این بخش، برخی از مهم‌ترین این رویکردها را بررسی می‌کنیم.

۱-۳-۱ LSTM کوانتومی و شبکه‌های عصبی بازگشتی کوانتومی

یکی از رویکردهای شهودی برای ورود محاسبات کوانتومی به پردازش زبان طبیعی، کوانتومی کردن معماری‌های موجود و مؤثر مانند حافظه طولانی کوتاه‌مدت است. ایده اصلی این است که لایه‌های کلاسیک شبکه عصبی را با مدارات کوانتومی پارامتری جایگزین کنیم. در این رویکرد، وضعیت پنهان^۱ و حافظه سلول^۲ به حالت‌های کوانتومی نگاشت می‌شوند، و دروازه‌های مختلف حافظه طولانی کوتاه‌مدت (دروازه فراموشی، ورودی، خروجی) با استفاده از دروازه‌های کوانتومی پیاده‌سازی می‌شوند.

در یک حافظه طولانی کوتاه‌مدت کوانتومی، هر گام زمانی شامل فرآیندهای زیر است: ابتدا ورودی کلاسیک (مثلاً بردار نمایش یک کلمه) به یک حالت کوانتومی کدگذاری می‌شود. سپس این حالت با حالت پنهان قبلی (که خود یک حالت کوانتومی است) ترکیب می‌شود. یک مدار کوانتومی پارامتری که شامل دروازه‌های چرخشی و درهم‌تنیدگی است، بر روی این حالت ترکیبی اعمال می‌شود تا حالت جدید را تولید کند. در نهایت، بخشی از این حالت اندازه‌گیری و به خروجی کلاسیک تبدیل می‌شود.

این رویکرد مزایایی دارد. از جمله اینکه می‌تواند از قدرت نمایشی فضای هیلبرت کوانتومی برای مدل‌سازی وابستگی‌های پیچیده زمانی بهره‌بردار. همچنین درهم‌تنیدگی کوانتومی می‌تواند روابط غیرخطی میان اجزای مختلف را به طور طبیعی نمایش دهد. با این حال، چالش‌های قابل توجهی نیز وجود دارد. نیاز به تبدیل مکرر میان فضای کلاسیک و کوانتومی در هر گام زمانی، هزینه محاسباتی قابل توجهی دارد و ممکن است منجر به از دست رفتن اطلاعات شود. علاوه بر این، برای دنباله‌های طولانی، نیاز به حفظ انسجام کوانتومی در طول چندین گام زمانی دشوار است. پیچیدگی مدار کوانتومی نیز به سرعت افزایش می‌یابد که با محدودیت‌های سخت‌افزاری فعلی سازگار نیست.

^۱ hidden state

^۲ cell state

۱-۳-۲ مکانیزم‌های توجه کوانتومی

رویکرد دیگر، پیاده‌سازی مکانیزم توجه^۱ با استفاده از الگوریتم‌های کوانتومی است. مکانیزم توجه که در معماری‌های مدرن مانند ترنسفورمر نقش کلیدی دارد، اساساً محاسبه میزان شباهت یا ارتباط میان بخش‌های مختلف ورودی است. در نسخه کوانتومی، می‌توان از آزمون تعویض کوانتومی^۲ برای محاسبه شباهت میان دو حالت کوانتومی استفاده کرد. همچنین تقویت دامنه یا تقویت دامنه^۳ می‌تواند برای تأکید بر بخش‌های مهم‌تر ورودی به کار رود. این رویکرد پتانسیل کاهش پیچیدگی محاسباتی را دارد، اما پیاده‌سازی عملی آن در سخت‌افزارهای فعلی چالش برانگیز است.

۱-۳-۳ مدل‌های نمایش کوانتومی برای کلمات

در این رویکرد، به جای استفاده از بردارهای کلاسیک برای نمایش کلمات (مانند Word2Vec)، هر کلمه به یک حالت کوانتومی نگاشت می‌شود. این نمایش‌های کوانتومی یا تعبیه‌های کوانتومی^۴ می‌توانند با استفاده از مدارات کوانتومی تغییراتی آموزشی ببینند. پارامترهای مدار به گونه‌ای بهینه می‌شوند که کلماتی که در زمینه‌های مشابه ظاهر می‌شوند حالت‌های کوانتومی مشابهی داشته باشند. شباهت میان کلمات می‌تواند با استفاده از معیارهای کوانتومی مانند فیدلیتی یا وفاداری^۵ اندازه‌گیری شود. این رویکرد می‌تواند از فضای نمایشی کوانتومی با ابعاد بالا بهره‌برد و احتمالاً روابط معنایی پیچیده‌تری را نسبت به روش‌های کلاسیک مدل کند.

¹ attention mechanism

² quantum swap test

³ amplitude amplification

⁴ quantum embeddings

⁵ fidelity

۱-۴-۴ رویکرد DisCoCat: معناشناسی ترکیبی-توزیعی-مقوله‌ای

۱-۴-۱ مبانی نظری و فلسفه رویکرد

رویکرد ترکیبی-توزیعی-مقوله‌ای^۱ که توسط کوک، سادرزاده و کلارک در سال ۲۰۱۰ معرفی شد [۹]، یکی از تأثیرگذارترین و پایه‌ای‌ترین چارچوب‌های ریاضی برای پردازش زبان طبیعی کوانتومی است. این رویکرد تلاش می‌کند تا پلی میان سه حوزه مهم ایجاد کند: معناشناسی توزیعی، دستور ترکیبی، و نظریه مقوله‌ها.

فلسفه اصلی ترکیبی-توزیعی-مقوله‌ای این است که معنای یک جمله از دو جنبه تشکیل می‌شود: اول، معنای کلمات فردی که از طریق رویکردهای توزیعی بر اساس زمینه استفاده آن‌ها به دست می‌آید. دوم، ساختار نحوی که نحوه ترکیب این معانی را مشخص می‌کند. نظریه مقوله‌ها زبان ریاضی مناسبی برای فرمال کردن این ترکیب فراهم می‌آورد.

در معناشناسی توزیعی^۲، معنای یک کلمه بر اساس زمینه‌هایی که در آن ظاهر می‌شود تعریف می‌شود. این ایده بر اساس فرضیه توزیعی است که قبلاً ذکر شد: کلماتی که در زمینه‌های مشابه ظاهر می‌شوند معانی مشابهی دارند. در مدل‌های توزیعی، معنای هر کلمه با یک بردار در فضای برداری نمایش داده می‌شود.

ساختار ترکیبی^۳ بر این اصل استوار است که معنای یک عبارت پیچیده، تابعی از معنای اجزای آن و نحوه ترکیب آن‌هاست. این اصل که به اصل ترکیب‌پذیری فرگه معروف است، یک مفروضه بنیادی در زبان‌شناسی است.

نظریه مقوله‌ها^۴ یک شاخه از ریاضیات است که با ساختارهای انتزاعی و تبدیلات میان آن‌ها سروکار دارد. در این نظریه، یک مقوله شامل اشیاء و مورفیس‌ها (نگاشت‌های) میان آن‌هاست. برای ترکیبی-توزیعی-مقوله‌ای، از نوع خاصی از مقوله‌ها به نام مقوله‌های فشرده متقارن یا مقوله‌های تکریختی متقارن^۵ استفاده می‌شود که دارای عملگر ضرب تانسوری \otimes هستند.

¹ DisCoCat (Distributional Compositional Categorical)

² distributional semantics

³ compositional structure

⁴ category theory

⁵ symmetric monoidal categories

۱-۴-۲ دستور پیش‌گروهی

در پایه رویکرد ترکیبی-توزیعی-مقوله‌ای، دستور پیش‌گروهی یا دستور پیش‌گروهی Pregroup Grammar [۱۰، ۱۱] قرار دارد که توسط لامبک توسعه یافته است. در این دستور، هر کلمه یک تایپ نحوی دارد که نحوه ترکیب آن با کلمات دیگر را مشخص می‌کند.

تایپ‌های پایه یا اتمی عبارتند از n برای اسم یا اسم^۱ و s برای جمله یا جمله^۲. از این تایپ‌های پایه می‌توان تایپ‌های پیچیده‌تر ساخت. برای هر تایپ A ، دو تایپ الحاقی یا الحاقی^۳ وجود دارد: A^r که الحاقی راست^۴ نامیده می‌شود و A^l که الحاقی چپ^۵ نامیده می‌شود. این تایپ‌های الحاقی نمایانگر این هستند که کلمه به چه نوع تایی در سمت راست یا چپ خود نیاز دارد. قوانین کاهش یا قوانین کاهش^۶ در دستور پیش‌گروهی بسیار ساده هستند:

$$n \otimes n^r \rightarrow 1 \quad \text{و} \quad n^l \otimes n \rightarrow 1 \quad (۴-۱)$$

در این رابطه، 1 نشان‌دهنده تایپ واحد یا خنثی است. این قوانین بیان می‌کنند که یک تایپ می‌تواند با تایپ الحاقی خود «حذف» یا «کاهش» شود.

به عنوان مثال، در زبان انگلیسی که ساختار فاعل-فعل-مفعول (SVO)^۷ دارد، یک فعل متعدی تایپ زیر را دارد:

$$n^r \otimes s \otimes n^l \quad (۵-۱)$$

این تایپ بیان می‌کند که فعل برای تشکیل جمله نیاز به یک اسم در سمت چپ (فاعل) و یک اسم در سمت راست (مفعول) دارد. وقتی این فعل با فاعل و مفعولش ترکیب می‌شود، داریم:

$$n \otimes (n^r \otimes s \otimes n^l) \otimes n \quad (۶-۱)$$

¹ noun

² sentence

³ adjoint

⁴ right adjoint

⁵ left adjoint

⁶ reduction rules

⁷ Subject-Verb-Object

با اعمال قوانین کاهش:

$$n \otimes (n^r \otimes s \otimes n^l) \otimes n \rightarrow (n \otimes n^r) \otimes s \otimes (n^l \otimes n) \rightarrow 1 \otimes s \otimes 1 \rightarrow s \quad (V-1)$$

که در نهایت به تایپ s (جمله) می‌رسد.

۱-۴-۳ نمایش دیاگرامی و دیاگرام‌های رشته‌ای

یکی از ویژگی‌های قدرتمند و جذاب رویکرد ترکیبی-توزیعی-مقوله‌ای، استفاده از دیاگرام‌های رشته‌ای^۱ است. این دیاگرام‌ها که در کتاب مرجع "تصویرسازی فرایندهای کوانتومی"^۲ [۱۲] به تفصیل توضیح داده شده‌اند، نمایشی بصری و شهودی از ساختار نحوی و معنایی جملات فراهم می‌کنند.

در دیاگرام‌های رشته‌ای، کلمات به صورت جعبه‌هایی نمایش داده می‌شوند که از آن‌ها سیم‌هایی (رشته‌ها) بیرون می‌آیند. این سیم‌ها نمایانگر تایپ‌های نحوی هستند. جهت عمودی دیاگرام نمایانگر ترکیب توالی^۳ است که با علامت \circ یا $>>$ نشان داده می‌شود. جهت افقی نمایانگر ضرب تانسوری^۴ است که با \otimes یا $@$ نمایش داده می‌شود.

عناصر اصلی در این دیاگرام‌ها عبارتند از: سیم^۵ که انتقال یک تایپ را نشان می‌دهد، فنجان^۶ که کاهش یا حذف دو تایپ الحاقی مجاور را نمایش می‌دهد، کلاهک^۷ که ایجاد یک جفت تایپ الحاقی را نشان می‌دهد، و تعویض^۸ که تعویض ترتیب دو سیم را نمایش می‌دهد.

برای مثال، جمله ساده انگلیسی "Alice loves Bob" را در نظر بگیرید. در این جمله، "Alice" و "Bob" هر کدام تایپ n (اسم) دارند، و "loves" تایپ $n^r \otimes s \otimes n^l$ دارد. دیاگرام این جمله نشان می‌دهد که چگونه تایپ‌های اسم‌ها با تایپ‌های الحاقی فعل کاهش می‌یابند تا در نهایت تایپ s (جمله) را تولید کنند. این نمایش بصری به فهم ساختار نحوی کمک می‌کند و همچنین راه را برای تبدیل به

¹ string diagrams

² Picturing Quantum Processes

³ sequential composition

⁴ tensor product

⁵ Wire

⁶ Cup

⁷ Cap

⁸ Swap

مدارات کوانتومی هموار می‌سازد.

مزیت اصلی دیاگرام‌های رشته‌ای این است که ارتباط مستقیمی با مدارات کوانتومی دارند. در واقع، دیاگرام‌های رشته‌ای و مدارات کوانتومی هر دو نمایش‌های بصری از مقوله‌های تک‌ریختی متقارن هستند، که این امر تبدیل میان آن‌ها را طبیعی و مستقیم می‌کند.

۱-۴-۴ نگاشت به فضای کوانتومی و تبدیل به مدار

یکی از شگفت‌انگیزترین جنبه‌های رویکرد ترکیبی-توزیعی-مقوله‌ای این است که به طور طبیعی و مستقیم با محاسبات کوانتومی سازگار است. این سازگاری ناشی از این واقعیت است که هم دستور پیش‌گرومی و هم محاسبات کوانتومی بر پایه ریاضیات مقوله‌های تک‌ریختی متقارن استوارند.

در این نگاشت، تایپ‌های نحوی به فضاهای هیلبرت نگاشت می‌شوند. به عنوان مثال، تایپ پایه n (اسم) به یک فضای هیلبرت مانند \mathbb{C}^2 (فضای دو کیوبیتی) نگاشت می‌شود، و تایپ s (جمله) نیز به یک فضای هیلبرت مشابه. ضرب تانسوری نحوی \otimes مستقیماً به ضرب تانسوری فضاهای هیلبرت نگاشت می‌شود. بنابراین اگر تایپ‌های A و B به فضاهای H_A و H_B نگاشت شوند، تایپ $A \otimes B$ به فضای $H_A \otimes H_B$ نگاشت می‌یابد.

کلمات که در ترکیبی-توزیعی-مقوله‌ای به عنوان بردارهای معنایی نمایش داده می‌شوند، در نسخه کوانتومی به حالت‌های کوانتومی یا عملگرهای کوانتومی تبدیل می‌شوند. برای مثال، یک اسم که در مدل توزیعی کلاسیک یک بردار در \mathbb{R}^d است، می‌تواند به یک حالت کوانتومی در فضای هیلبرت مربوطه تبدیل شود. یک فعل که تایپ پیچیده‌تری دارد، به یک عملگر کوانتومی یا یک مدار کوانتومی پارامتری نگاشت می‌شود.

عناصر ساختاری دیاگرام‌های رشته‌ای نیز معادل‌های کوانتومی دارند. عملیات فنجان که دو تایپ الحاقی را کاهش می‌دهد، در فضای کوانتومی به حالت‌های درهم‌تنیده مانند حالت بل^۱ و عملیات اندازه‌گیری نگاشت می‌شود. عملیات تعویض که ترتیب دو سیم را تعویض می‌کند، به دروازه تعویض کوانتومی^۲ نگاشت می‌یابد. سیم‌ها که تایپ‌ها را منتقل می‌کنند، به عملگر همانی^۳ در فضای کوانتومی

^۱ Bell state

^۲ SWAP gate

^۳ Identity

نگاشت می‌شوند.

این نگاشت طبیعی امکان تبدیل مستقیم یک دیاگرام نحوی ترکیبی-توزیعی-مقوله‌ای به یک مدار کوانتومی را فراهم می‌آورد. این فرآیند یک از مراحل کلیدی در پردازش زبان طبیعی کوانتومی است، زیرا پل میان تحلیل نحوی زبانی و اجرای کوانتومی را ایجاد می‌کند.

۱-۴-۵ کتابخانه لامبک: ابزار کامل برای پردازش زبان طبیعی کوانتومی

کتابخانه Lambeq [۱۳] یک کتابخانه پایتون منبع باز و جامع است که توسط تیم کوانتینیوم^۱ توسعه یافته و پیاده‌سازی کامل رویکرد ترکیبی-توزیعی-مقوله‌ای برای پردازش زبان طبیعی کوانتومی را فراهم می‌کند. این کتابخانه طراحی شده تا تمام مراحل پایپلاین^۲ از جمله تبدیل متن به مدار کوانتومی، آموزش مدل، و ارزیابی را پوشش دهد.

معماری لامبک شامل چندین مؤلفه اصلی است که به صورت ماژولار طراحی شده‌اند. اولین مؤلفه، تحلیل‌گر^۳ است که وظیفه تبدیل جملات متنی به دیاگرام‌های نحوی را بر عهده دارد. برای زبان انگلیسی، BobcatParser ارائه شده که بر اساس گرامر ترکیبی مقوله‌ای (CCG)^۴ عمل می‌کند. این تحلیل‌گر جملات را تجزیه کرده و به دیاگرام‌های ترکیبی-توزیعی-مقوله‌ای تبدیل می‌کند. علاوه بر این، تحلیل‌گرهای دیگری نیز مانند SpacyTokeniser و BagsOfWordsParser برای کاربردهای خاص وجود دارند.

مؤلفه دوم، بازنویس^۵ است که وظیفه ساده‌سازی و بهینه‌سازی دیاگرام‌ها را دارد. دیاگرام‌های نحوی اولیه ممکن است پیچیده باشند و شامل عناصری مانند فنجان‌های متعدد یا ساختارهای اضافی. بازنویس‌ها این دیاگرام‌ها را ساده می‌کنند تا برای تبدیل به مدار کوانتومی مناسب‌تر باشند. مهم‌ترین بازنویس، RemoveCupsRewriter است که عملیات‌های فنجان را حذف می‌کند و دیاگرام را به شکل خطی‌تری در می‌آورد.

سومین مؤلفه کلیدی، آنساز است که دیاگرام‌های ساده‌شده را به مدارات کوانتومی پارامتری تبدیل

¹ Quantinuum

² pipeline

³ Parser

⁴ Combinatory Categorical Grammar

⁵ Rewriter

می‌کند. لامبک چندین آنساز مختلف ارائه می‌دهد، از جمله IQPAnsatz^۱، Sim14Ansatz، و MPSAnsatz^۲. هر آنساز استراتژی متفاوتی برای نگاشت دیاگرام به مدار دارد و برای کاربردهای خاص بهینه شده است. IQPAnsatz به ویژه در کاربردهای پردازش زبان طبیعی کوانتومی محبوب است زیرا توازن خوبی میان قدرت نمایی و پیچیدگی مدار دارد.

مؤلفه آموزش دهنده^۳، وظیفه آموزش پارامترهای مدار کوانتومی را بر عهده دارد. لامبک از الگوریتم‌های بهینه‌سازی مختلفی پشتیبانی می‌کند، از جمله QuantumTrainer که برای آموزش بر روی سخت‌افزار کوانتومی طراحی شده، و PytorchTrainer و JAXTrainer که برای شبیه‌سازی کلاسیک کارآمد استفاده می‌شوند. این آموزش دهنده‌ها از توابع هزینه مختلفی مانند Binary Cross-Entropy و Mean Squared Error پشتیبانی می‌کنند.

در نهایت، پسانه^۴، محیطی را برای اجرای مدارات کوانتومی فراهم می‌کند. لامبک با چندین پسانه سازگار است، از جمله شبیه‌سازهایی مانند NumpyBackend برای آزمایش سریع، و پسانه‌های کوانتومی واقعی مانند pytket که به پلتفرم‌های مختلف کوانتومی از جمله IBM Quantum، Quantinuum، و IonQ متصل می‌شود. این انعطاف‌پذیری امکان آزمایش و توسعه بر روی شبیه‌ساز و سپس اجرا بر روی سخت‌افزار واقعی را فراهم می‌آورد.

۱-۴-۶ تحقیقات مرتبط و کاربردها

از زمان معرفی رویکرد DisCoCat، تحقیقات گسترده‌ای در زمینه کاربردهای آن انجام شده است. یکی از اولین و موفق‌ترین کاربردها، طبقه‌بندی متن و دسته‌بندی جملات بوده است. لورنز و همکاران در مقاله تأثیرگذار خود با عنوان "QNLP in Practice" [۱۴] نشان دادند که مدل‌های مبتنی بر DisCoCat می‌توانند در وظایف دسته‌بندی جملات کوتاه نتایج رقابتی با روش‌های کلاسیک داشته باشند، حتی با مجموعه داده‌های آموزشی کوچک‌تر. آن‌ها مدل خود را بر روی سخت‌افزار کوانتومی واقعی اجرا کردند و نشان دادند که رویکرد کوانتومی در عمل قابل پیاده‌سازی است.

تحلیل احساسات^۵ نیز یکی دیگر از کاربردهای مورد بررسی بوده است. در این وظیفه، هدف تعیین

¹ Instantaneous Quantum Polynomial

² Matrix Product State

³ Trainer

⁴ Backend

⁵ sentiment analysis

احساس مثبت یا منفی یک جمله است. رویکرد DisCoCat می‌تواند ساختار نحوی جمله را در نظر بگیرد و تأثیر کلمات منفی‌ساز یا تشدیدکننده را به طور صحیح مدل کند. برای مثال، تفاوت میان "فیلم خوب نبود" و "فیلم بد نبود" که از نظر لغوی مشابه هستند اما معانی متضادی دارند، می‌تواند با تحلیل صحیح ساختار نحوی تشخیص داده شود.

پاسخ به پرسش^۱ و استنتاج متنی^۲ نیز در چند مطالعه بررسی شده‌اند. در این وظایف، مدل باید رابطه منطقی میان دو جمله را تعیین کند. دیاگرام‌های DisCoCat می‌توانند ساختار معنایی هر دو جمله را نمایش دهند و مقایسه آن‌ها را تسهیل کنند.

مزایای اصلی رویکرد DisCoCat نسبت به رویکردهای دیگر در چندین جنبه نمایان است. اول، قابلیت تفسیر بالا: دیاگرام‌های رشته‌ای قابل مشاهده و درک هستند و می‌توان به طور بصری ساختار نحوی و جریان معنا را دنبال کرد. این ویژگی در مقایسه با مدل‌های جعبه سیاه مانند شبکه‌های عصبی عمیق بسیار ارزشمند است. دوم، پایه ریاضی محکم: استفاده از نظریه مقوله‌ها چارچوب دقیق و رسمی برای استدلال در مورد معنا فراهم می‌کند. سوم، مدل‌سازی صریح ترکیب‌پذیری: برخلاف روش‌هایی که ساختار نحوی را به صورت ضمنی یاد می‌گیرند، DisCoCat به طور مستقیم از دستور زبان استفاده می‌کند. چهارم، کارایی داده^۳: به دلیل استفاده از ساختار نحوی، این روش معمولاً به داده‌های آموزشی کمتری نسبت به مدل‌های یادگیری عمیق نیاز دارد.

۵-۱ پردازش زبان فارسی و QNLP

۱-۵-۱ اهمیت و انگیزه تحقیق

زبان فارسی یکی از زبان‌های کهن و غنی جهان است که توسط بیش از ۱۱۰ میلیون نفر در ایران، افغانستان، تاجیکستان و جوامع ایرانی‌تبار سرتاسر جهان صحبت می‌شود. علی‌رغم اهمیت این زبان، منابع و ابزارهای پردازش زبان طبیعی فارسی به ویژه در حوزه‌های نوین مانند QNLP به طور قابل توجهی محدودتر از زبان انگلیسی است. این شکاف فناوری، انگیزه قوی برای توسعه سیستم‌های QNLP برای زبان فارسی فراهم می‌کند.

^۱ question answering

^۲ textual entailment

^۳ data efficiency

توسعه ابزارهای پردازش زبان فارسی کاربردهای عملی فراوانی دارد که به زندگی روزمره میلیون‌ها کاربر فارسی‌زبان مرتبط است. موتورهای جستجوی فارسی می‌توانند از درک بهتر پرسش‌های کاربران و محتوای صفحات وب فارسی بهره ببرند. ترجمه ماشینی فارسی-انگلیسی و بالعکس، که نقش مهمی در ارتباطات بین‌المللی و دسترسی به دانش جهانی دارد، می‌تواند از روش‌های پیشرفته‌تر سود ببرد. تحلیل احساسات در شبکه‌های اجتماعی فارسی برای درک افکار عمومی و بازخورد مشتریان اهمیت دارد. خلاصه‌سازی خودکار اخبار و اسناد فارسی می‌تواند به مدیریت بهتر سیل اطلاعات کمک کند. دستیارهای صوتی و چت‌بات‌های فارسی‌زبان می‌توانند خدمات بهتری به کاربران ارائه دهند. سیستم‌های پرسش و پاسخ فارسی می‌توانند دسترسی به اطلاعات را تسهیل کنند.

از منظر تحقیقاتی، توسعه QNLP برای زبان فارسی چالش‌های منحصر به فردی را ارائه می‌دهد که حل آن‌ها می‌تواند به پیشرفت عمومی این حوزه کمک کند. تفاوت‌های ساختاری میان فارسی و انگلیسی، به ویژه در ترتیب کلمات و وجود عناصر نحوی خاص، نیازمند تطبیق و گاه بازتعریف رویکردهای موجود است. موفقیت در این زمینه نشان می‌دهد که رویکردهای QNLP قابلیت تعمیم به زبان‌های متنوع را دارند و محدود به زبان انگلیسی نیستند.

۱-۵-۲ ویژگی‌های ساختاری و نحوی زبان فارسی

زبان فارسی دارای ویژگی‌های منحصر به فردی است که آن را از زبان انگلیسی متمایز می‌کند و چالش‌هایی برای پردازش خودکار ایجاد می‌نماید.

از نظر الفبا و نگارش، فارسی از الفبای عربی-فارسی استفاده می‌کند که در آن چهار حرف اضافی ("پ"، "چ"، "ژ"، "گ") نسبت به الفبای عربی وجود دارد. نوشتار فارسی از راست به چپ (RTL)^۱ است که برای نمایش و پردازش الگوریتمی نیازمند ابزارهای خاص است. حروف فارسی اشکال مختلفی بسته به موقعیت در کلمه (ابتدا، وسط، انتها یا منفرد) دارند که این موضوع توکنیزیشن در سطح کاراکتر را پیچیده می‌کند. علاوه بر این، مصوت‌های کوتاه (حروف کوتاه) معمولاً نوشته نمی‌شوند، که باعث ابهام در خواندن و پردازش می‌شود.

از نظر واژگانی، فارسی دارای سیستم غنی ساخت‌واژه است. کلمات مرکب در فارسی بسیار رایج هستند، مانند "نرم‌افزار"، "هوش مصنوعی"، "یادگیری ماشین" و "اشکال‌زدایی". این کلمات از دو یا

¹ Right-to-Left

چند جزء تشکیل شده‌اند که گاه با نیم‌فاصله و گاه بدون آن نوشته می‌شوند، که این امر مشکلاتی در توکنیزیشن ایجاد می‌کند. افعال مرکب نیز بسیار متداول هستند، که از یک جزء اسمی یا صفتی به همراه یک فعل کمکی (معمولاً "کردن"، "شدن" یا "داشتن") تشکیل می‌شوند، مانند "اجرا کردن"، "تصمیم گرفتن"، و "استفاده کردن". یک ویژگی مهم دیگر، اضافه کشیده است که با صدای " (کسره) نمایش داده می‌شود اما معمولاً نوشته نمی‌شود. این عنصر نحوی برای اتصال اسم به صفت، اسم به اسم، یا اسم به مکمل به کار می‌رود، مانند "کتاب جالب" (کتاب [ا] جالب).

از نظر ساختار نحوی، مهم‌ترین تفاوت فارسی با انگلیسی، ترتیب اجزای جمله است. فارسی یک زبان SOV^۱ است، در حالی که انگلیسی SVO است. بنابراین جمله‌ای مانند "مرد غذا را خورد" در انگلیسی به صورت "The man ate the food" می‌آید که ترتیب مفعول و فعل متفاوت است. این تفاوت تأثیر مستقیم بر تایپ‌گذاری دستور پیش‌گروهی دارد.

یکی از ویژگی‌های برجسته فارسی، وجود نشانگر مفعول "را" است که بعد از مفعول مستقیم معین می‌آید. این عنصر نحوی در انگلیسی وجود ندارد و تشخیص و مدل‌سازی آن در رویکرد DisCoCat نیازمند تعریف تایپ نحوی جدید است. فارسی همچنین اجازه حذف ضمائر فاعلی را می‌دهد (زبان pro-drop)، بنابراین جملاتی مانند "رفتم" (من رفتم) بدون ذکر صریح فاعل کامل هستند. ترتیب صفت و موصوف نیز با انگلیسی متفاوت است: در فارسی صفت بعد از موصوف می‌آید ("کتاب بزرگ") در حالی که در انگلیسی قبل از آن است "big book".

۳-۵-۱ چالش‌های پردازش زبان فارسی و QNLP

پردازش خودکار زبان فارسی با چالش‌های متعددی مواجه است که برخی از آن‌ها عمومی و برخی دیگر خاص رویکرد کوانتومی هستند.

یکی از اساسی‌ترین چالش‌ها، ابهام در توکنیزیشن یا تقسیم متن به کلمات است. در فارسی، مرز میان کلمات همیشه با فاصله^۲ مشخص نمی‌شود. برخی کلمات مرکب با نیم‌فاصله نوشته می‌شوند، برخی بدون هیچ جداکننده‌ای، و برخی با فاصله کامل. این عدم یکنواختی باعث می‌شود تشخیص اینکه آیا دو کلمه مجاور باید به عنوان یک واحد یا دو واحد جداگانه در نظر گرفته شوند دشوار باشد. علاوه بر این،

^۱ Subject-Object-Verb

^۲ Space

پیشوندها و پسوندها معمولاً به کلمه اصلی چسبیده‌اند و تشخیص آن‌ها نیازمند تحلیل ریخت‌شناسی است.

تحلیل ریخت‌شناسی فارسی پیچیدگی‌های خاص خود را دارد. فعل‌های فارسی بر اساس زمان (گذشته، حال، آینده)، شخص (اول، دوم، سوم) و شمار (مفرد، جمع) تصریف می‌شوند، که منجر به شکل‌های متعدد برای هر فعل می‌شود. برخی اسم‌ها جمع‌های غیرقیاسی دارند که باید به صورت واژگانی یاد گرفته شوند. سیستم ساخت‌واژه غنی فارسی امکان ایجاد کلمات جدید با ترکیب پیشوندها، پسوندها و عناصر مختلف را فراهم می‌کند که مدل‌سازی آن دشوار است.

کمبود منابع داده‌ای نیز یک مشکل جدی است. پیکره‌های زبانی فارسی که با برچسب‌های نحوی، نقش‌های معنایی یا سایر اطلاعات زبان‌شناختی حاشیه‌نویسی شده باشند، به طور قابل توجهی کوچک‌تر از معادل‌های انگلیسی هستند. این محدودیت بر توانایی آموزش مدل‌های داده‌محور تأثیر می‌گذارد. همچنین ابزارهای پردازش زبان فارسی نسبت به انگلیسی کمتر توسعه یافته و استاندارد شده هستند.

تنوع املائی در متون فارسی، به ویژه در محتوای تولید شده توسط کاربران در وب و شبکه‌های اجتماعی، چالش دیگری است. استانداردهای املائی همیشه رعایت نمی‌شوند، و تنوع در نوشتن برخی حروف مانند "ی" و "ی" یا "ک" و "ک" رایج است. مشکلات کدگذاری کاراکتر^۱ نیز می‌تواند بروز کند، به ویژه تفاوت میان UTF-8 و فرم‌های نمایشی عربی.

برای QNLP فارسی، چالش‌های خاصی وجود دارد. اولاً، عدم وجود یک پارسر^۲ کوانتومی آماده برای فارسی. BobcatParser که در لامبک ارائه شده تنها برای زبان انگلیسی است و نمی‌تواند مستقیماً برای فارسی استفاده شود. بنابراین نیاز به توسعه یک پارسر سفارشی برای فارسی وجود دارد، که کاری وقت‌گیر و نیازمند تخصص زبان‌شناختی است.

ثانیاً، تفاوت در ساختار نحوی نیاز به بازتعریف تاپ‌های دستور پیش‌گروهی دارد. در زبان انگلیسی SVO، فعل متعدی تاپ $n^r \otimes s \otimes n^l$ دارد. اما در فارسی SOV، این تاپ باید تطبیق یابد. علاوه بر این، وجود نشانگر مفعول "را" نیازمند تعریف یک تاپ نحوی جدید است که در دستور پیش‌گروهی استاندارد وجود ندارد. این تطبیقات نه تنها نیازمند درک عمیق دستور پیش‌گروهی، بلکه نیازمند درک دقیق ساختار نحوی فارسی نیز هستند.

^۱ character encoding

^۲ Parser

ثالثاً، نمایش دوجهتی و نمایش دیاگرام‌ها برای متون RTL چالش‌برانگیز است. ابزارهای موجود برای رسم دیاگرام‌های رشته‌ای معمولاً برای متون LTR طراحی شده‌اند و نمایش صحیح فارسی نیازمند تنظیمات خاص یا توسعه ابزارهای جدید است. کتابخانه‌هایی مانند arabic-reshaper و python-bidi ممکن است برای این منظور لازم باشند.

۱-۵-۴ ابزارها و منابع موجود برای پردازش زبان فارسی

علی‌رغم چالش‌های ذکر شده، چند ابزار و منبع مفید برای پردازش زبان فارسی در دسترس هستند که می‌توانند در توسعه سیستم‌های QNLP فارسی مورد استفاده قرار گیرند.

مهم‌ترین کتابخانه پردازش زبان فارسی، هضم^۱ [۱۵] است که یک کتابخانه پایتون جامع و منبع‌باز برای NLP فارسی محسوب می‌شود. هضم مجموعه‌ای از ابزارهای مفید را ارائه می‌دهد که شامل نرمال‌ساز^۲ برای یکنواخت‌سازی متن و حذف تنوع‌های املائی، توکنیزر کلمه^۳ برای جداسازی متن به کلمات، توکنیزر جمله^۴ برای تقسیم متن به جملات، برچسب‌زن اجزای کلام یا POS Tagger که بر اساس مدل‌های آماری یا یادگیری ماشین کلمات را برچسب‌گذاری می‌کند، ریشه‌یاب^۵ برای یافتن ریشه کلمات، و تحلیل‌گر وابستگی نحوی^۶ برای استخراج ساختار نحوی جملات است. این ابزارها پایه و اساس بسیاری از سیستم‌های پردازش زبان فارسی هستند و در این تحقیق نیز مورد استفاده قرار گرفته‌اند. علاوه بر هضم، ابزارهای دیگری نیز وجود دارند. ParsiNLU یک مجموعه معیار (benchmark) و پایگاه داده برای وظایف مختلف NLP فارسی ارائه می‌دهد که برای ارزیابی عملکرد مدل‌ها مفید است. ابزارهای مختلف دیگری نیز توسط محققان ایرانی در دانشگاه‌ها و مراکز تحقیقاتی توسعه یافته‌اند.

از نظر منابع داده، چند پیکره مهم در دسترس هستند. پیکره بیجنخان^۷ یکی از اولین و معتبرترین پیکره‌های برچسب‌گذاری شده فارسی است که حاوی متون برچسب‌گذاری شده با اجزای کلام است. درخت وابستگی فارسی^۸ [۱۶] که بر اساس چارچوب وابستگی جهانی^۹ ساخته شده، حاوی جملات

^۱ Hazm

^۲ Normalizer

^۳ Word Tokenizer

^۴ Sentence Tokenizer

^۵ Lemmatizer

^۶ Dependency Parser

^۷ Bijankhan Corpus

^۸ Persian Dependency Treebank

^۹ Universal Dependencies

تجزیه‌شده نحوی است و برای آموزش پارسر ها بسیار ارزشمند است. همچنین دیتاست هایی برای وظایف خاص مانند شناسایی موجودیت‌های اسمی (NER)^۱ در دسترس هستند.

۱-۵-۵ اهداف و چشم‌انداز این تحقیق

هدف اصلی این تحقیق، توسعه یک سیستم پردازش زبان طبیعی کوانتومی برای زبان فارسی بر اساس رویکرد DisCoCat و با استفاده از کتابخانه لامبک است. این هدف کلی به اهداف فرعی مشخصی تقسیم می‌شود.

اول، طراحی و پیاده‌سازی یک پارسر سفارشی فارسی به نام PersianCatParser است که مشابه BobcatParser برای انگلیسی عمل کند. این parser باید بتواند جملات فارسی را به دیاگرام‌های DisCoCat تبدیل کند. برای این منظور، لازم است تایپ‌های دستور پیش‌گروهی مناسب برای ساختار SOV فارسی تعریف شوند. همچنین باید مکانیزمی برای مدیریت نشانگر مفعول "را" و سایر ویژگی‌های خاص فارسی مانند کلمات مرکب، افعال مرکب، و صفات پیاده‌سازی شود.

دوم، یکپارچه‌سازی با کتابخانه هضم برای بهره‌برداری از ابزارهای موجود پردازش زبان فارسی است. از POS Tagger هضم برای برچسب‌گذاری خودکار اجزای کلام استفاده می‌شود. نرمال‌سازی و پیش‌پردازش متن فارسی با استفاده از ابزارهای استاندارد هضم انجام می‌گیرد. همچنین از قابلیت‌های هضم برای شناسایی کلمات مرکب و افعال مرکب بهره‌برداری می‌شود.

سوم، آزمایش سیستم بر روی یک وظیفه واقعی NLP برای ارزیابی عملکرد آن است. وظیفه انتخابی، دسته‌بندی جملات فارسی به دو دسته فناوری اطلاعات یا IT و غذا یا Food است. این وظیفه مشابه آزمایش‌های انجام شده برای انگلیسی است و امکان مقایسه را فراهم می‌کند. نتایج با روش‌های پایه^۲ های کلاسیک مقایسه می‌شوند و تحلیل جامعی از نتایج، محدودیت‌ها و فرصت‌های بهبود ارائه می‌شود. چهارم، این تحقیق شامل توسعه دو نسخه از سیستم است که به صورت تکاملی پیشرفت کرده‌اند. نسخه V00 یک پیاده‌سازی ساده با قوانین دستی و تایپ‌گذاری ثابت است که برای درک مفاهیم پایه و آزمایش ایده‌های اولیه طراحی شده است. نسخه V01 پیشرفته‌تر است و از کتابخانه هضم برای پردازش زبانی استفاده می‌کند، قابلیت‌های بیشتری برای مدیریت ساختارهای پیچیده‌تر دارد، و کیفیت بهتری در

¹ Named Entity Recognition

² baseline

تشخیص نقش‌های نحوی و پارس ارائه می‌دهد.

مشارکت‌های اصلی این تحقیق عبارتند از: ارائه پارسر کوانتومی برای زبان فارسی بر اساس رویکرد DisCoCat، تطبیق دستور پیش‌گروهی با ساختار نحوی SOV فارسی و ویژگی‌های خاص این زبان، ایجاد یک دیتاست برچسب‌گذاری‌شده فارسی برای وظیفه دسته‌بندی که می‌تواند برای تحقیقات آینده مفید باشد، نشان دادن امکان‌پذیری و قابلیت تعمیم رویکرد DisCoCat به زبان‌های غیرانگلیسی با ساختارهای نحوی متفاوت، و ارائه کدهای منبع باز و مستندسازی شده که می‌تواند مبنایی برای تحقیقات بعدی در این حوزه باشد.

۶-۱ جمع‌بندی

در این فصل، سیری جامع از تاریخچه پردازش زبان طبیعی تا مرزهای جدید پردازش زبان طبیعی کوانتومی داشتیم. ابتدا تحولات پردازش زبان طبیعی را از دهه‌های میانی قرن بیستم تا امروز بررسی کردیم، از روش‌های مبتنی بر قاعده گرفته تا مدل‌های نمایش توزیعی کلمات، شبکه‌های عصبی بازگشتی، معماری ترنسفورمر، و در نهایت مدل‌های زبانی بزرگی که امروزه توانایی‌های شگفت‌انگیزی در فهم و تولید زبان از خود نشان می‌دهند.

سپس به رویکرد کوانتومی در پردازش زبان طبیعی پرداختیم. مبانی محاسبات کوانتومی شامل برهم‌نهی، درهم‌تنیدگی و تداخل را توضیح دادیم و اشاره کردیم که چگونه این ویژگی‌های منحصر به فرد می‌توانند برای مدل‌سازی زبان به کار گرفته شوند. مزایای بالقوه رویکرد کوانتومی از جمله نمایش طبیعی ترکیب‌پذیری و توانایی مدل‌سازی همبستگی‌های غیرکلاسیک را بررسی کردیم، و همچنین چالش‌های جدی آن در عصر NISQ فعلی را مورد بحث قرار دادیم.

رویکرد DisCoCat به عنوان یکی از قدرتمندترین و پایه‌ای‌ترین چارچوب‌های ریاضی برای پردازش زبان طبیعی کوانتومی معرفی شد. این رویکرد با ترکیب معناشناسی توزیعی، دستور پیش‌گروهی، و نظریه مقوله‌ها، پلی میان زبان‌شناسی و فیزیک کوانتومی ایجاد می‌کند. دیاگرام‌های رشته‌ای نمایشی بصری و شهودی از ساختار نحوی و معنایی جملات فراهم می‌کنند و به طور طبیعی به مدارات کوانتومی نگاشت می‌شوند. کتابخانه لامبک با ارائه ابزارهای کامل از پارسر گرفته تا پسانه‌های کوانتومی، پیاده‌سازی عملی این رویکرد را ممکن ساخته است.

در نهایت، به چالش‌های خاص و فرصت‌های پردازش زبان فارسی پرداختیم. ویژگی‌های منحصر به فرد زبان فارسی از جمله ساختار SOV، نشانگر مفعول "را"، کلمات و افعال مرکب، و اضافه کشیده را تشریح کردیم. چالش‌های فنی مانند توکنیزیشن، کمبود منابع، و عدم وجود پارسر کوانتومی آماده برای فارسی را مورد بحث قرار دادیم. اهداف این تحقیق برای ایجاد یک سیستم QNLP فارسی و مشارکت‌های مورد انتظار آن را تشریح نمودیم.

در فصل‌های بعدی، به جزئیات فنی خواهیم پرداخت. فصل دوم ابزارها و مفاهیم فنی پردازش زبان را بررسی می‌کند. فصل سوم به تفصیل به رویکرد DisCoCat و دیاگرام‌های نحوی برای جملات فارسی می‌پردازد. فصل چهارم پیاده‌سازی، آموزش مدل، نتایج آزمایش‌ها و چشم‌انداز آینده را ارائه خواهد کرد.

Bibliography

- [1] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [2] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- [3] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- [4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.
- [5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.
- [6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.
- [7] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd. Quantum machine learning. *Nature*, 549:195–202, 2017.
- [8] M. Schuld, I. Sinayskiy, and F. Petruccione. An introduction to quantum machine learning. *Contemporary Physics*, 56(2):172–185, 2015.

- [9] B. Coecke, M. Sadrzadeh, and S. Clark. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, 36:345–384, 2010.
- [10] J. Lambek. *Type Grammar Revisited*. Springer, 1999.
- [11] A. Preller and J. Lambek. Free compact 2-categories. *Mathematical Structures in Computer Science*, 17(2):309–340, 2007.
- [12] B. Coecke and A. Kissinger. *Picturing Quantum Processes: A First Course in Quantum Theory and Diagrammatic Reasoning*. Cambridge University Press, 2017.
- [13] D. Kartsaklis, I. Fan, R. Yeung, A. Pearson, R. Lorenz, A. Toumi, G. de Felice, K. Meichanetzidis, S. Clark, and B. Coecke. lambeq: An efficient high-level python library for quantum nlp. *arXiv preprint arXiv:2110.04236*, 2021.
- [14] R. Lorenz, A. Pearson, K. Meichanetzidis, et al. Qnlp in practice: Running compositional models of meaning on a quantum computer. *Journal of Artificial Intelligence Research*, 76:1305–1342, 2023.
- [15] A. Aleahmad et al. Hazm: Python library for persian natural language processing. *GitHub Repository*, 2016. <https://github.com/sobhe/hazm>.
- [16] M. Seraji and J. Nivre. A universal dependencies treebank for persian. *Proceedings of LREC*, 2016.
- [17] K. Meichanetzidis, A. Toumi, G. de Felice, and B. Coecke. Foundations for near-term quantum natural language processing. *arXiv preprint arXiv:2012.03755*, 2020.
- [18] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe. Circuit-centric quantum classifiers. *Physical Review A*, 101(3):032308, 2020.
- [19] B. Sabeti, P. Hosseini, S. Ghassemi, and M. Ghassemi. Hazm: A persian natural language processing toolkit. *GitHub Repository*, 2022. <https://github.com/roshan-research/hazm>.
- [20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

-
- [21] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567:209–212, 2019.
 - [22] M. Shamsfard. Challenges and open problems in persian text processing. In *Proceedings of LTC*, 2010.
 - [23] D. Contributors. Discopy: The python toolkit for computing with monoidal categories, 2025. Accessed: 2025-10-19.
 - [24] V. Bergholm, J. Izaac, M. Schuld, et al. PennyLane: Automatic differentiation of hybrid quantum-classical computations. *arXiv preprint arXiv:1811.04968*, 2018.
 - [25] G. Aleksandrowicz, T. Alexander, P. Barkoutsos, et al. Qiskit: An open-source framework for quantum computing. *Zenodo*, 2019.
 - [26] N. Wiebe and K. Meichanetzidis. Practical quantum natural language processing. *arXiv preprint arXiv:2401.12345*, 2024.
 - [27] A. Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O’Reilly Media, 3rd edition, 2023.
 - [28] S. Raschka. *Build a Large Language Model (From Scratch)*. Manning Publications, 2024.