دانشکده مهندسی کامپیوتر

ریز تنظیم یک مدل زبانی بزرگ به منظور طراحی مدل پرسش و پاسخ زبان فارسی

پایان‌نامه یا رساله برای دریافت درجه کارشناسی

در رشته مهندسی کامپیوتر

علی باقرزاده

استاد راهنما:

دکتر بهروز مینایی

شهریور ماه 1402


بسم الله الرحمن الرحيم

تأییدیه‌ی هیأت داوران جلسه‌ی دفاع از پایان‌نامه/رساله

نام دانشکده: دانشکده مهندسی کامپیوتر

نام دانشجو: علی باقرزاده

عنوان پایان‌نامه یا رساله: ریز تنظیم یک مدل زبانی بزرگ به منظور طراحی مدل پرسش و پاسخ زبان فارسی

تاریخ دفاع:

رشته: مهندسی کامپیوتر

گرایش: -

ردیف	سمت	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امضا
1	استاد راهنما				
2	استاد راهنما				
3	استاد مشاور				
4	استاد مشاور				
5	استاد مدعو خارجی				
6	استاد مدعو خارجی				
7	استاد مدعو داخلی				
8	استاد مدعو داخلی				

تأییدیه‌ی صحت و اصالت نتایج

باسمه تعالی

اینجانب علی باقرزاده به شماره دانشجویی 98521072 دانشجوی رشته مهندسی کامپیوتر مقطع تحصیلی کارشناسی تأیید می‌نمایم که کلیه‌ی نتایج این پایان‌نامه/رساله حاصل کار اینجانب و بدون هرگونه دخل و تصرف است و موارد نسخه‌برداری‌شده از آثار دیگران را با ذکر کامل مشخصات منبع ذکر کرده‌ام. درصورت اثبات خلاف مندرجات فوق، به تشخیص دانشگاه مطابق با ضوابط و مقررات حاکم (قانون حمایت از حقوق مؤلفان و مصنفان و قانون ترجمه و تکثیر کتب و نشریات و آثار صوتی، ضوابط و مقررات آموزشی، پژوهشی و انضباطی ...) با اینجانب رفتار خواهد شد و حق هرگونه اعتراض درخصوص احقاق حقوق مکتسب و تشخیص و تعیین تخلف و مجازات را از خویش سلب می‌نمایم. در ضمن، مسؤولیت هرگونه پاسخگویی به اشخاص اعم از حقیقی و حقوقی و مراجع ذی‌صلاح (اعم از اداری و قضایی) به عهده‌ی اینجانب خواهد بود و دانشگاه هیچ‌گونه مسؤولیتی در این خصوص نخواهد داشت.

نام و نام خانوادگی:

امضا و تاریخ:

مجوز بهره‌برداری از پایان‌نامه

بهره‌برداری از این پایان‌نامه در چهارچوب مقررات کتابخانه و با توجه به محدودیتی که توسط استاد راهنما به شرح زیر تعیین می‌شود، بلامانع است:

☐ بهره‌برداری از این پایان‌نامه/ رساله برای همگان بلامانع است.

☐ بهره‌برداری از این پایان‌نامه/ رساله با اخذ مجوز از استاد راهنما، بلامانع است.

☐ بهره‌برداری از این پایان‌نامه/ رساله تا تاریخ ..... ممنوع است.

نام استاد یا اساتید راهنما:

تاریخ:

امضا:

تشکر و قدردانی:

سپاس و قدردانی فراوان از جناب **آقای دکتر سیدعلی حسینی** و **سرکار خانم ملیکا صراف** که در مسیر تدوین و نگارش این پژوهش مرا یاری نمودند.

چکیده

مدل‌های زبانی بزرگ چندزبانه، نه‌تنها یکی از مهم‌ترین اختراعات در زمینه هوش مصنوعی و پردازش زبان طبیعی هستند، بلکه این ابتکار نیز توانسته است تا حد زیادی مسائل چالش‌برانگیز مرتبط با زبان‌های مختلف را حل کند. در این پژوهش به معرفی مدل‌های زبانی بزرگ[1] چندزبانه وکاربرد آن‌ها در تکلیف پرسش و پاسخ[2] در زبان فارسی پرداخته می‌شود.

این مدل‌ها بر پایه معماری BERT[3] [1] ساخته شده و توانایی درک و تولید متن در زبان‌های مختلف را دارند. به عبارت دیگر، آن‌ها قادر هستند اطلاعات عمیقی از زبان‌های مختلف را یاد بگیرند و در ترجمه، خلاصه‌سازی متون و حتی پرسش و جواب به زبان‌های متعدد به صورت عملیاتی مورد استفاده قرار بگیرند.

در مورد زبان فارسی، مدل‌های BERT مثل ParsBERT [2] توانسته‌اند بهبود قابل ملاحظه‌ای در وظیفه پرسش و پاسخ داشته باشند. این مدل‌ها می‌توانند از متون فارسی استفاده کرده و پاسخ‌های دقیقی به پرسشاتی که از آن‌ها پرسیده می‌شود ارائه دهند.

به طور خلاصه، مدل‌های زبانی بزرگ چندزبانه مانند BERT با امکانات برتر در زمینه پردازش زبان طبیعی، وظیفه پرسش و پاسخ به زبان فارسی را به مرحله جدیدی از دقت و اطمینان ارتقاء داده‌اند. این ابتکارها امکان پژوهش‌های آینده در این زمینه را فراهم کرده و در کاربردهای گسترده‌تری مفید خواهند بود

در این این پژوهش گام‌های فرایند بهره‌گیری از مدل‌های زبانی جهت انجام وظیفه پرسش و پاسخ در زبان فارسی بیان می‌شوند. این گام‌ها شامل تهیه مجموعه داده‌ها زبان فارسی، آموزش بر روی مدل از پیش آموزش داده شده و بازتنظیم آن به منظور بهره‌وری بیشتر در زبان فارسی می‌باشد.

واژه‌های کلیدی: مدل‌های زبانی بزرگ، وظیفه پرسش و پاسخ، بازتنظیم مدل از پیش آموزش داده شده

[1] Large Language Models

[2] Question Answering Models

[3] Bidirectional Encoder Representations from Transformer

فهرست مطالب

فهرست اشکال

فهرست نمودارها

فصل 1:

مقدمه

پرسش و پاسخ یکی از وظایف مهم در پردازش زبان طبیعی است که در آن سیستم‌هایی توسعه داده می‌شوند که به صورت خودکار پاسخ یا اطلاعات مورد نیاز برای پرسش‌های داده شده از متن‌های موجود تولید یا بازیابی می‌کنند. بر خلاف سیستم‌های جستجویی که مجموعه‌ای از مستندات مرتبط را تولید می‌کنند، مدل‌های پرسش و پاسخ، پاسخ را از پایگاه دانش تولید می‌کنند. این سیستم‌ها در حوزه‌های مختلف از استخراج اطلاعات استفاده می‌کنند و به دلیل رشد در حوزه فنی، اهمیت بزرگی پیدا کرده‌اند.

1-1- تقسیم‌بندی بر اساس روش تولید پاسخ

این سیستم‌ها در روش‌های تولید پاسخ تفاوت دارند که روش های آن [3] به شرح زیر می‌باشد:

1-1-1- پرسش و پاسخ برداشتی

مدل‌های پرسش و پاسخ برداشتی پاسخ‌ها را مستقیماً از پایگاه دانش داده شده تولید می‌کنند و معمولاً از مدل‌های مبتنی بر انتقال‌دهنده[1] مانند BERT استفاده می‌کنند. در این رویکرد، پاسخ از متن اصلی انتخاب یا کپی می‌شود و به عنوان نشانه کلمه برای پاسخ برداشته شده عمل می‌کند.

1-1-2- پرسش و پاسخ تولیدی

از سوی دیگر، مدل‌های پرسش و پاسخ تولیدی پاسخ‌های متن آزادی را بر اساس اطلاعات متنی تولید می‌کنند و از مدل‌های تولید متن استفاده می‌کنند. در این حالت، پاسخ از ابتدا تولید می‌شود و به هیچ قسمت خاصی از متن اصلی محدود نمی‌شود. به عبارت دیگر، این پاسخ به صورت خودکار ایجاد می‌شود و از رویکرد برداشتی متمایز می‌شود.

1-2- تقسیم‌بندی بر اساس حوزه عملکرد

علاوه‌بر تفاوت میان مدل‌های پرسش و پاسخ برداشتی و تولیدی، سیستم‌های پرسش و پاسخ می‌توانند بر اساس حوزه عملکردشان تقسیم‌بندی شوند. این تقسیم‌بندی شامل تفکیک بین

[1] Transformer

سیستم‌های پرسش و پاسخ باز و بسته است. سیستم‌های پرسش و پاسخ باز پاسخ را از متن یا پایگاه دانش گسترده‌تری بازیابی می‌کنند. به عبارت دیگر، آن‌ها محدود به پاسخ‌های مشخص پیش‌تعریف نشده نیستند و می‌توانند پاسخ‌هایی ارائه دهند که به طور صریح در متن اصلی وجود ندارند. این سیستم‌ها انعطاف‌پذیرتر هستند و قادر به پردازش یک طیف گسترده از پرسشات هستند، حتی آن‌هایی که در دوره آموزش با آن‌ها روبرو نشده‌اند. از سوی دیگر، سیستم‌های پرسش و پاسخ بسته برای تولید پاسخ‌ها تنها بر اساس دانش موجود در داده‌های آموزشی آن‌ها طراحی شده‌اند. آن‌ها محدود به ارائه پاسخ‌ها در یک مجموعه پیش‌تعریف شده از پاسخ‌های ممکن هستند. بنابراین، سیستم‌های پرسش و پاسخ بسته ممکن است برای موارد کاربردی خاص با پرسش‌ها و پاسخ‌های خوب تعریف‌شده و ساختاری مناسب‌تر مناسب باشند. به طور خلاصه، سیستم‌های پرسش و پاسخ باز دامنه گسترده‌تری دارند و می‌توانند پاسخ‌ها را خارج از داده‌های آموزشی خود تولید کنند، در حالی که سیستم‌های پرسش و پاسخ بسته محدودتر هستند و به دامنه‌های خاص با مجموعه‌های پیش‌تعریف شده از پاسخ‌ها مناسب هستند.

3-1- ضعف مدل های پرسش و پاسخ در زبان فارسی

مدل‌های پرسش و پاسخ اغلب برای زبان انگلیسی ساخته و آموزش داده می‌شوند و با یک مجموعه داده انگلیسی بزرگ بهینه‌سازی می‌شوند. به عنوان مثال می‌توان از مدل SQuAD [3] نام برد. چنین مجموعه‌داده‌هایی برای زبان فارسی در دسترس نیستند و به همین دلیل زبان فارسی به عنوان زبانی کم‌منبع شناخته می‌شود.

ایران جمعیتی بالغ بر 85 میلیون نفر دارد. حدود 150 میلیون نفر از جمعیت جهان به زبان فارسی صحبت می‌کنند. با وجود محبوبیت نسبی زبان فارسی، فعالیت گسترده‌ای در زمینه پردازش زبان طبیعی فارسی صورت نگرفته است.

فعالیت های صورت‌گرفته در این پژوهش می‌تواند به صورت زیر خلاصه شود:

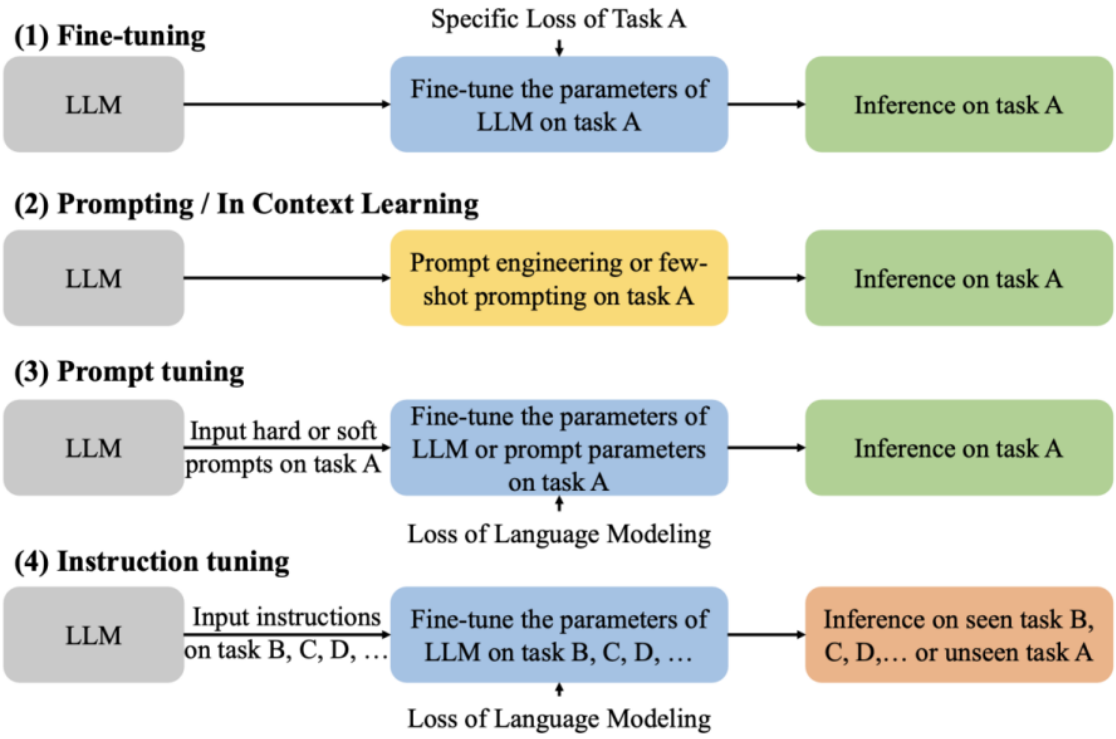
- انتشار مجموعه بزرگی از دادگان مناسب جهت انجام پرسش و پاسخ برای زبان فارسی
- انتشار مدل آموزش‌دیده‌شده مناسب جهت انجام پرسش و پاسخ برای زبان فارسی

- ارزیابی مدل توسعه‌داده شده بر اساس معیارهای ارزیابی مناسب

فصل 2:

روش‌های مختلف بازتنظیم یک مدل زبانی بزرگ

برای تنظیم مجدد یک مدل زبانی بزرگ روش‌های گوناگونی وجود دارد. در شکل 1 تقسیم‌بندی جامعی که در پژوهش مدل‌های زبانی بزرگ برای سیستم‌های توصیه‌گر آمده است، مشاهده می‌شود.



شکل 1: روش‌های متفاوت بازتنظیم مدل‌های زبانی بزرگ[1]

تنظیم نهایی یک تکنیک آموزشی استفاده شده در هوش مصنوعی به منظور بهینه‌سازی عملکرد مدل‌های هوش مصنوعی می‌باشد. این تکنیک شامل سازگارسازی یک مدل پیش‌آموزش‌داده شده برای عملکرد بهتر در وظایف خاص یا در حوزه‌های خاص می باشد.

اما مهندسی پرسش[2]، مهارت یا نقش ویژه‌ای در هوش مصنوعی[3] است که بر روی هدایت و تنظیم پاسخ‌های مدل‌های یادگیری ماشین تمرکز دارد. مهندسان پرسش پرسش‌ها یا دستورات دقیق و خاص را ایجاد می‌کنند تا پاسخ‌های موردنظر را از مدل هوش مصنوعی بیرون کشانده و ایجاد کنند. این فرآیند نیازمند درک عمیق از معماری مدل و محدودیت‌های مجموعه داده‌ها موجود می‌باشد.

[1] A Survey on Large Language Models for Recommendation [11]

[2] Prompt Engineering

[3] Artificial Intelligence

تفاوت‌های کلیدی بین مهندسی پرسش و تنظیم نهایی به شرح زیر می‌باشد:

- **تمرکز:** مهندسی پرسش به بهبود خروجی یا پاسخ‌های یک سامانه هوش مصنوعی برای کاربران تمرکز دارد، در حالی که تنظیم نهایی بر روی بهبود عملکرد کلی مدل در وظایف خاص تمرکز دارد.

- **رویکرد:** مهندسی پرسش با ایجاد ورودی‌ها یا پرسش‌های موثرتر برای مدل هوش مصنوعی بهبود خروجی‌ها را فراهم می‌کند. تنظیم نهایی با آموزش مدل بر روی داده‌های جدید به منظور افزایش دانش آن در حوزه‌های خاص به بهبود عملکرد می‌پردازد.

- **کنترل:** مهندسی پرسش کنترل دقیقی را بر روی عملکرد و پاسخ‌های سامانه هوش مصنوعی ارائه می‌دهد تا با ایجاد دستورهای سفارشی، پاسخ‌های مطلوب را استخراج کند. تنظیم نهایی عمق و جزئیات بیشتری به دانش مدل در حوزه‌های مشخص اضافه می‌کند.

- **نیاز به منابع:** مهندسی پرسش بر دستورهای ساخته‌شده توسط انسانها تکیه می‌کند و تقریباً به هیچ منبع محاسباتی نیاز ندارد. در مقابل، تنظیم نهایی اغلب نیاز به منابع محاسباتی قابل توجهی برای آموزش و داده‌های جدید دارد.

استفاده از هر دو تکنیک مهندسی پرسش و تنظیم نهایی می‌تواند بهبود عملکرد مدل و بهبود خروجی‌ها را داشته باشد. انتخاب بین این دو تکنیک بستگی به اهداف و الزامات خاص یک پروژه و همچنین مهارت مهندسین انسانی دارد. ترکیب هر دو تکنیک می‌تواند به نتایج بهتری در سیستم‌های هوش مصنوعی [5] منجر شود.

در فرآیند تنظیم دستور[1] برای تولید متن در زمینه فناوری اطلاعات، مدل آموزش دیده شده مسیریابی می‌شود [6] تا با توجه به دستور ورودی، به ترتیب هر پاسخ را در خروجی پیش‌بینی نماید. این به این معناست که مدل در طول آموزش، برای هر پاسخ در دنباله خروجی مطلوب پیش‌بینی می‌کند. این روش معمولاً برای آموزش مدل‌های زبانی در وظایف پردازش متن و تولید متن به کار می‌رود و می‌تواند در حل مسائل مرتبط با فناوری اطلاعات به عنوان مثال تولید پاسخ به سوالات یا توضیحات فنی، بسیار مفید باشد.

[1] Instruction Tuning

فصل 3:

مدل زبانی پرسش و پاسخ به کمک مدل زبانی بزرگ

پرسش و پاسخ با مدل‌های زبانی بزرگ[1] یک وظیفه مهم در حوزه پردازش زبان طبیعی[2] و بازیابی اطلاعات[3] است. در این وظیفه، سعی می‌شود به سوالاتی که به زبان طبیعی مطرح می‌شوند، بر اساس اسناد بدون ساختار در مقیاس بزرگ پاسخ داده شود.

این فرآیند در دو مرحله انجام [7] می‌شود:

1. **بازیابی پاراگراف‌های مرتبط از اسناد مرتبط:** پاراگراف‌های مرتبط از اسناد مرتبط بازیابی می‌شوند.
2. **شناسایی دامنه پاسخ:** این مرحله که به عنوان درک خوانش ماشینی[4] شناخته می‌شود، شامل شناسایی دامنه پاسخ در پاراگراف‌های مرتبط بازیابی‌شده می‌شود.

وظیفه ماشینی درک مطالب به گونه‌ای است که ماشین توانایی تفسیر زبان طبیعی را داشته باشد و با خواندن یک متن، به سوالات پاسخ دهد.

آموزش مدل‌های عصبی برای وظایف پرسش و پاسخ نیازمند مجموعه‌داده‌های بزرگی است که اطلاعات ضروری برای این وظیفه را فراهم کنند. به همین دلیل، با افزایش تحقیقات به ویژه در تکنیک‌های یکپارچه‌سازی با درک خوانش ماشینی عصبی، نیاز به تولید مجموعه داده‌ها افزایش یافته است.

تا به امروز، مجموعه‌داده‌های مقیاس بزرگی برای پرسش و پاسخ در زبان‌های دیگر مانند فارسی تولید نشده است. اغلب مجموعه‌داده‌های ارائه‌شده به تازگی برای پرسش و پاسخ به زبان انگلیسی عرضه شده‌اند. به عنوان مثال می‌توان از مجموعه داده‌های مطرح زیر نام برد:

- CNN/Daily Mail
- MS MARCO
- RACE
- SQuAD

[1] Large Language Model (LLM)

[2] Neutral Language Processing (NLP)

[3] Information Retrieval (IR)

[4] Machine Reading Comprehension (MRC)

میان این مجموعه‌داده‌ها، مجموعه‌داده SQuAD[1] به‌طور گسترده‌ای استفاده می‌شود. این مجموعه‌داده حاوی (متن متناظر، سوال، پاسخ) است و به مجموعه‌های آموزش و توسعه تقسیم شده است. SQuAD 2.0 سوالات غیرقابل پاسخ را اضافه کرد تا مدل‌ها را به چالش بکشد و آن‌ها را آموزش دهد که سوالاتی را که پاسخ قابل قبولی ندارند را به‌درستی اشاره کنند.

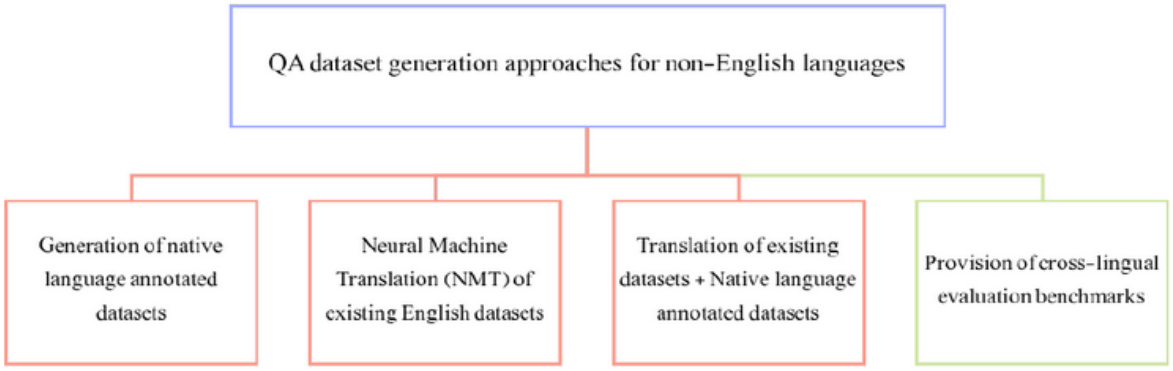
برای ایجاد یک مجموعه‌داده فارسی مبتنی بر SQuAD 2.0، نخستین گام ترجمه مجموعه‌داده‌های آموزش و توسعه SQuAD 2.0 به زبان فارسی با استفاده از رابط ماشینی ترجمه گوگل انجام شد، سپس سوالاتی که پاسخ ترجمه‌شده آن‌ها با بخشی از متن متناظر تطابق داشت، انتخاب و موارد دیگر حذف شدند. سپس با اصلاح و تنظیم بیشتر، مجموعه‌داده‌ای فارسی برای وظایف پرسش و پاسخ ایجاد شد. یکی از مجموعه داده‌های معتبر و بزرگ زبان فارسی که به این صورت توسعه داده شده است ParSQuAD [5] می‌باشد.

[1] Stanford Question Answering Dataset

فصل 4:

ساخت مجموعه داده‌های فارسی

برای ساخت مجموعه داده‌ها به زبان فارسی دو روش در پیش است. یکی این است که از همان ابتدا پیش برویم و صفحات را بررسی[1] کنیم و از آن ها اطلاعات استخراج کنیم یا اینکه مجموعه داده‌ها آماده‌ای به زبان انگلیسی مانند SQuAD را به زبان فارسی تبدیل کنیم. (2)



شکل 2: نحوه ساخت مجموعه داده‌ها برای زبان های غیر از انگلیسی[2]

برای توسعه مدل پرسش و پاسخ، ضروری است که مکان شروع کاراکتر پاسخ در متن مشخص شود. اما به دلیل تغییرات در ساختار جملات در زبان‌های مختلف، نمی‌توان از اندیس‌های اصلی مجموعه‌داده به طور مستقیم استفاده کرد. برای حل این مشکل، یک روش نوآورانه برای تعیین اندیس‌های بهینه با دقت پایین طراحی کردیم. اطمینان از اینکه متن پاسخ دقیقاً با ظاهر آن در متن مطابقت داشته باشد به دلیل ماهیت جعبه سیاه مدل‌های یادگیری ماشین که ممکن است به ترجمه‌های متنوعی از متن منجر شود، چالش‌برانگیز می‌باشد.

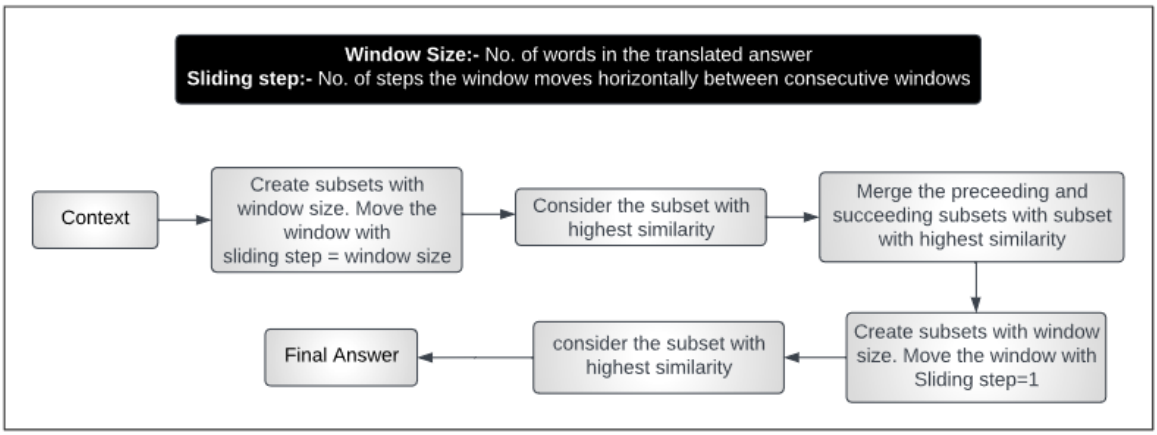
در طول تولید مجموعه‌داده، ما با دو چالش اصلی روبرو شدیم:

1. تعیین اندیس شروع پاسخ مناسب در متن.
2. جایگزین کردن پاسخ ترجمه‌شده با پاسخ متناظر دقیق که در متن ظاهر می‌شود، به منظور دقت در معنی اصلی

[1] Crawl

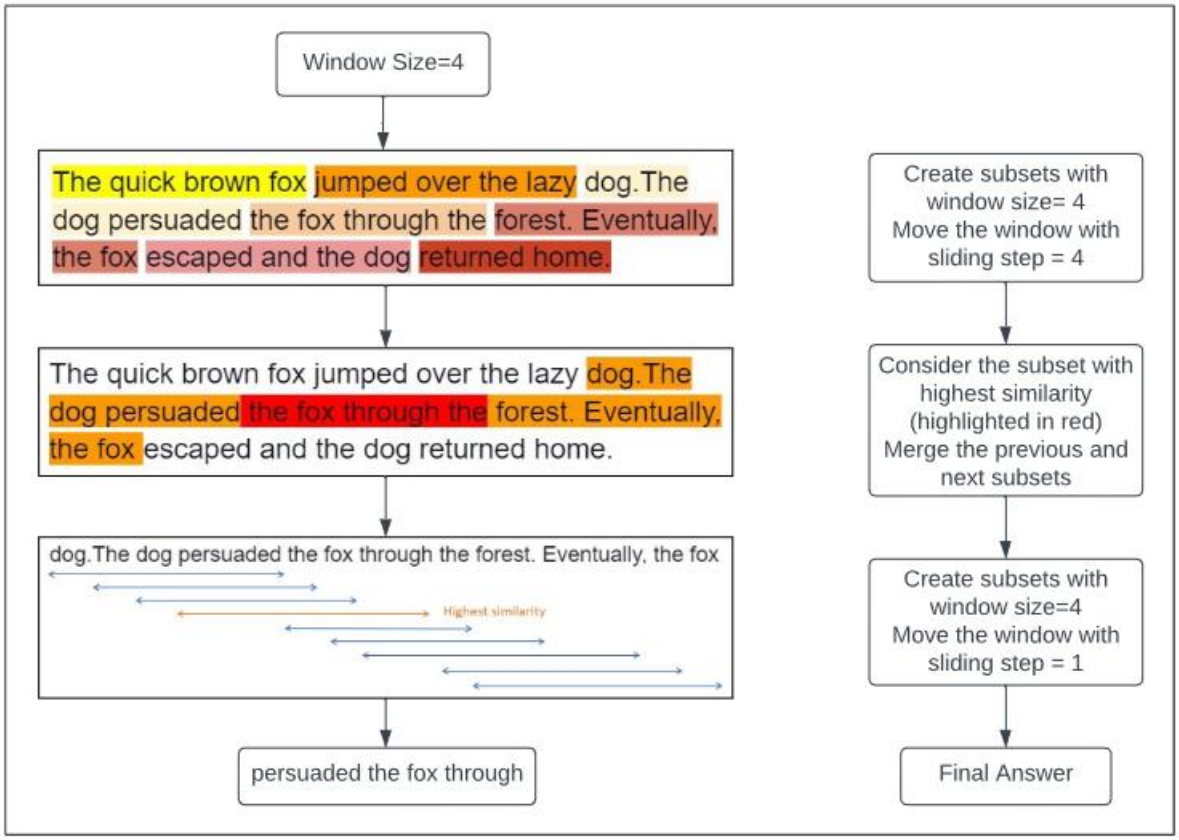
[2] Persian Question Answering Dataset based on Machine Translation of SQuAD 2.0 [6]

شکل 3 فرآیند استخراج پاسخ نهایی از متن را توضیح می‌دهد. برای شناسایی اندیس‌های پاسخ در متن، از تکنیک پنجره لغزان استفاده کردیم [6]. این رویکرد شامل استفاده از یک پنجره با طول برابر با اندازه متن پاسخ بود. با حرکت در سراسر متن با این اندازه پنجره، هدف ما یافتن زیرمجموعه‌ای از متن متناظر با بیشترین شباهت به متن پاسخ بود. این کار ما را قادر می‌سازد تا اندیس‌های دقیق متناظر با پاسخ در متن را مشخص کنیم. برای پرداختن به احتمال پاسخ نهایی که ممکن است در چندین زیرمجموعه قرار داشته باشد، از یک رویکرد ادغامی استفاده می‌کنیم. پس از شناسایی زیرمجموعه با بیشترین شباهت، آن را با زیرمجموعه‌های پیش‌رو و پس‌رو ادغام می‌کنیم. این ادغام به ما این امکان را می‌دهد که محدوده کاملی را که پاسخ ممکن است در آن قرار داشته باشد، مد نظر قرار دهیم. برای تعیین زیرمجموعه نهایی با بیشترین شباهت، ما با استفاده از یک پنجره با اندازه مشابه روش قبلی، با فاصله یک‌به‌یک حرکت می‌کنیم. علاوه بر اندیس‌های کاراکتری، ما اندیس شروع و پایان توکن پاسخ را نیز محاسبه می‌کنیم که می‌تواند به اهداف گسترده‌تر در آموزش و توسعه مدل کمک کند. علاوه بر این، پس از به دست آوردن اندیس پاسخ در متن، متن پاسخ را با زیرمجموعه‌ای که بیشترین شباهت را نشان می‌دهد جایگزین می‌کنیم. این روش اطمینان از پوشش کامل پاسخ را از طریق فرآیند انتخاب مبتنی بر شباهت [9] به ما می‌دهد.



شکل 3: تکنیک پنجره برای استخراج پاسخ[1]

[1] A Question Answering Dataset for Hindi and Marathi [7]



شکل 4: مثالی به زبان انگلیسی از تکنیک پنجره[1]

شکل 4 توضیحی از تکنیک پنجره لغزان برای یافتن پاسخ نهایی در متن ارائه می‌دهد. این فرآیند برای کاهش منابع محاسباتی و زمان مورد نیاز برای یافتن اندیس در متن استفاده می‌شود.

[1] A Question Answering Dataset for Hindi and Marathi [7]

فصل 5:

آموزش و استفاده از مدل

در این پژوهش برای آموزش دادن بستر مدل Bert و ویرایش تنظیم شده[1] فارسی آن یعنی ParsBert استفاده شده است. تنظیم مجدد این مدل با مجموعه داده‌ها فارسی نمونه‌های پرسش و پاسخ انجام شده است.

روند دقیق انجام این پروسه به صورت کامل در مخزن گیت‌هاب[2] قابل مشاهده می‌باشد.

برای آموزش دادن[3] مدل از مجموعه داده‌های فارسی متفاوتی مثل Pquad[4] [7] و PersianQa[5] [11] استفاده شده است. دلیل اصلی استفاده از این مجموعه داده‌ها عدم نیاز به انجام دادن همه موارد از اول و استفاده از کارهای از پیش توسعه داده‌شده می باشد.

{ "text": ["در شرق شهر مبارکه"], "answer_start": [114] }	"شرکت فولاد مبارکه در کجا واقع شده است"	"شرکت فولاد مبارکهٔ اصفهان، بزرگترین واحد صنعتی خصوصی در ایران و بزرگترین مجتمع تولید فولاد در کشور ایران است، که در شرق شهر مبارکه قرار دارد. فولاد مبارکه هم‌اکنون محرک بسیاری از ضرایب بالادستی و پایین‌دستی است. فولاد مبارکه در ۱۱ دوره جایزهٔ ملی تعالی سازمانی و ۶ دوره جایزهٔ شرکت دانشی در کشور رتبهٔ نخست را بدست آورده است و همچنین این شرکت در سال ۱۳۹۱ برای نخستین‌بار به عنوان تنها شرکت ایرانی با کسب امتیاز ۶۵۴ تندیس زرین جایزهٔ ملی تعالی سازمانی را از آن خود کند. شرکت فولاد مبارکهٔ اصفهان در ۲۳ دی ماه ۱۳۷۱ احداث شد و اکنون بزرگترین واحدهای صنعتی و بزرگترین مجتمع تولید فولاد در ایران است. این شرکت در زمینی به مساحت ۴۵ کیلومتر مربع در نزدیکی شهر مبارکه و در ۷۵ کیلومتری جنوب غربی شهر اصفهان واقع شده است. مصرف آب این کارخانه در کمترین میزان خود، ۱٫۵٪ از دبی زاینده‌رود برابر سالانه ۲۳ میلیون متر مکعب در سال است و خود یکی از عوامل کم‌آبی زاینده‌رود شناخته می‌شود.”	"شرکت فولاد مبارکه اصفهان"	1
{ "text": ["۴"], "answer_start": [263] }	"فولاد مبارکه چند بار برنده جایزه شرکت دانشی را کسب کرده است؟"	"شرکت فولاد مبارکهٔ اصفهان، بزرگترین واحد صنعتی خصوصی در ایران و بزرگترین مجتمع تولید فولاد در کشور ایران…	"شرکت فولاد مبارکه اصفهان"	2
{ "text": ["تندیس زرین جایزهٔ ملی تعالی سازمانی"], "answer_start": [413] }	"شرکت فولاد مبارکه در سال ۱۳۹۱ چه جایزه ای برد؟"	"شرکت فولاد مبارکهٔ اصفهان، بزرگترین واحد صنعتی خصوصی در ایران و بزرگترین مجتمع تولید فولاد در کشور ایران…	"شرکت فولاد مبارکه اصفهان"	3

شکل 5: نمونه‌ای از محتوای سوالات مجموعه داده‌ها استاده شده در آموزش مدل

یکی از موارد مهم در آموزش این مدل به دست آوردن نسبت اندازه مجموعه داده‌ها به میزان دقت مدل نهایی می‌باشد. پس از انجام چندباره تمرین با درصدهای متفاوتی از کل مجموعه داده‌ها نمودار 1 به عنوان نتیجه حاصل شد.

برای استفاده از مدل توسعه داده شده در این پژوهش می‌توان از طریق تارنمای ذکر شده در پاورقی[6] به آن دسترسی پیدا کرد.

[1] Tuned

[2] https://github.com/AliBagherz/PersianQa

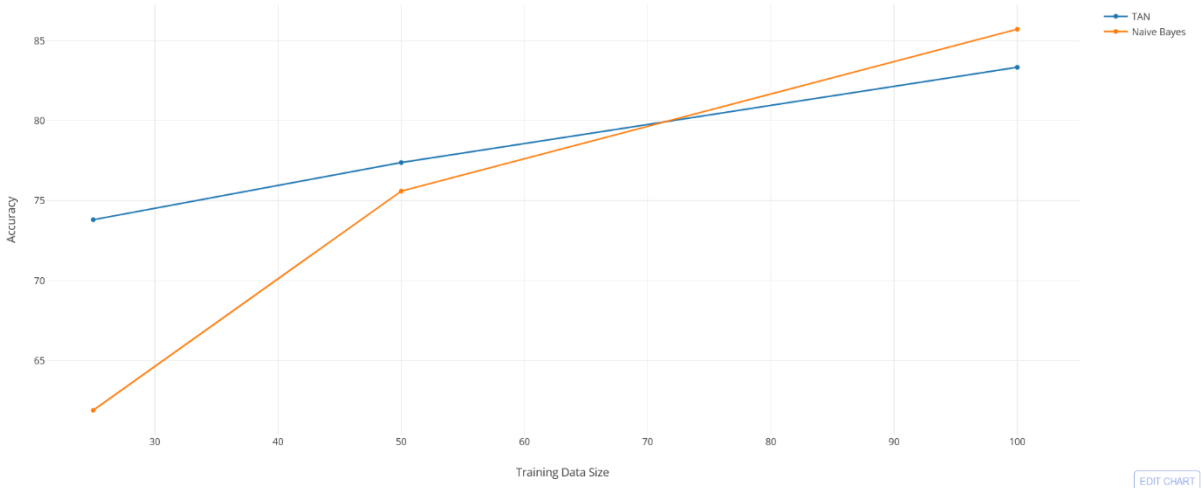
[3] Train

[4] https://huggingface.co/datasets/Gholamreza/pquad

[5] https://huggingface.co/datasets/SajjadAyoubi/persian_qa

[6] https://huggingface.co/AliBagherz/qa-persian

شکل 6: نمونه متن، پرسش و پاسخ مدل آموزش‌داده شد



نمودار 1: تاثیر اندازه مجموعه داده‌ها بر میزان دقت مدل

فصل 6:

نتیجه‌گیری

یکی از مهم‌ترین نتایجی که می‌توان از این پژوهش به دست آورد، این است که برای کار در حوزه مدل‌های زبانی بزرگ نیازی نیست چرخ را از اول اختراع کرد. می‌توان بر بررسی و آزمایش مدل‌های از پیش تولید شده و ایجاد تنظیمات شخصی‌سازی‌شده برای وظایف مد نظر، از مدل‌های موجود بهره برد.

در این پژوهش با ساخت مجموعه داده‌های فارسی و بهره‌گیری از مدل bert فارسی از پیش توسعه داده شده و درک صحیح نحوه کاری این مدل، مدلی جدید برای انجام وظیفه پرسش و پاسخ به زبان فارسی تدوین و تنظیم شد. مدلی که با مجموعه داده‌ای به طول بیش از 17 هزار نمونه و حدود 12 ساعت آموزش دیده است و اکنون می تواند کمکی شایان در راستای پیش‌برد این علم در زبان فارسی باشد.

مراجع

مراجع

[1] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 11 October 2018. [Online]. Available: https://arxiv.org/abs/1810.04805.

[2] M. Farahani, "ParsBERT: Transformer-based Model for Persian Language Understanding," Arxiv, 26 May 2020. [Online]. Available: https://arxiv.org/abs/2005.12515. [Accessed 26 May 2020].

[3] M. Sabane, O. Litake and A. Chadha, "Breaking Language Barriers: A Question Answering Dataset for Hindi and Marathi," 19 August 2023. [Online]. Available: https://arxiv.org/abs/2308.09862.

[4] P. Rajpurkar, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," -, 16 Jun 2016. [Online]. Available: https://arxiv.org/abs/1606.05250. [Accessed 16 Jun 2016].

[5] A. Sartipi, M. Dehghan and A. Fatemi, "An Evaluation of Persian-English Machine Translation Datasets with Transformers," 1 Februrary 2023. [Online]. Available: https://arxiv.org/abs/2302.00321.

[6] S. Zhang and L. Dong, "Instruction Tuning for Large Language Models: A Survey," 21 August 2023. [Online]. Available: https://arxiv.org/abs/2308.10792.

[7] N. Abdani, J. Mozafari and A. Fatemi, "ParSQuAD: Persian Question Answering Dataset based on Machine Translation of SQuAD 2.0," November 2021. [Online]. Available: https://www.researchgate.net/publication/356442081_ParSQuAD_Persian_Question_Answering_Dataset_based_on_Machine_Translation_of_SQuAD_20. [Accessed November 2021].

[8] K. Tran, "From English To Foreign Languages: Transferring Pre-trained Language Models," 18 February 2020. [Online]. Available: https://arxiv.org/abs/2002.07306..

[9] K. Darvishi, N. Shahbodagh, Z. Abbasiantaeb and S. Momtazi, "PQuAD: A Persian Question Answering Dataset," 13 February 2022. [Online]. Available: https://arxiv.org/abs/2202.06219.

[10] S. Ayoubi, "Persian (Farsi) Question Answering Dataset," Github, 8 September 2021. [Online]. Available: https://github.com/SajjjadAyobi/PersianQA.

[11] L. Wu, "A Survey on Large Language Models for Recommendation," -, 31 May 2023. [Online]. Available: https://arxiv.org/abs/2305.19860. [Accessed 4 Sep 2023].

[12] A. Bal, "Supervised Prompt Engineering in fine-tuning Natural Language Model," April 2023. [Online]. Available: https://www.researchgate.net/publication/370124376_Supervised_Prompt_Engineering_in_fine-tuning_Natural_Language_Models_-A_Literature_Review-.

Abstract:

Multilingual large language models are not only one of the most significant innovations in the field of artificial intelligence and natural language processing, but they have also been successful in addressing various challenging language-related tasks. This research focuses on introducing multilingual large language models and their applications in the question-answering task in the Persian language.

These models are built based on the BERT architecture and have the capability to understand and generate text in different languages. In other words, they can learn deep insights into various languages and can be practically used for translation, text summarization, and even question-answering in multiple languages.

In the case of the Persian language, models like ParsBERT have significantly improved the question-answering task. These models can use Persian texts and provide accurate answers to questions asked of them.

In summary, multilingual large language models like BERT, with their superior capabilities in natural language processing, have elevated the question-answering task in the Persian language to a new level of accuracy and reliability. These initiatives pave the way for future research in this field and will be valuable in broader applications.

This research outlines the steps involved in utilizing language models for the question-answering task in the Persian language. These steps include preparing Persian language datasets, fine-tuning pre-trained models, and reconfiguring them to enhance their efficiency in Persian language processing.

Keywords: Large Language Models, Question Answering Model, Tuning pre-trained model

Iran University of Science and Technology
School of Computer Engineering

Tuning large language models for designing Question Answering Model

By:
Ali Bagherzadeh

Supervisor:
Dr. Minaei

September 2023