دانشکده علوم ریاضی

پروژه نهایی دوره کارشناسی

عنوان

خلاصه سازی متن مقالات علمی

نگارش

حنانه اسفندیار

علی‌رضا عباس‌زاده

استاد راهنما

دکتر مجتبی روحانی

شهریور ماه ۱۴۰۳

چکیده

این مطالعه به بررسی و توسعه روش‌های خلاصه‌سازی متن با تمرکز بر مقالات علمی می‌پردازد و از مدل‌های ترانسفورمر به‌عنوان ابزار اصلی در این فرآیند استفاده می‌کند. با توجه به حجم بالای اطلاعات علمی و نیاز به استخراج سریع و مؤثر اطلاعات کلیدی، این مطالعه به تحلیل دو نوع خلاصه‌سازی، یعنی استخراجی و انتزاعی، می‌پردازد. در این راستا، ابتدا مجموعه داده‌های متنوعی از مقالات علمی جمع‌آوری و پیش‌پردازش شده و سپس مدل‌های مختلف ترانسفورمر برای تولید خلاصه‌های دقیق و کارآمد به کار گرفته می‌شوند. نتایج حاصل از ارزیابی‌های کمی و کیفی نشان‌دهنده بهبود قابل توجه در دقت و کیفیت خلاصه‌ها نسبت به روش‌های سنتی است. این پژوهش نه‌تنها به درک بهتر فرآیند خلاصه‌سازی کمک می‌کند، بلکه راهکارهای نوینی برای تسهیل دسترسی به اطلاعات علمی ارائه می‌دهد و به‌عنوان مبنایی برای تحقیقات آینده در این حوزه محسوب می‌شود.

واژه های کلیدی: خلاصه‌سازی متن، مدل‌های ترانسفورمر، پردازش زبان طبیعی، یادگیری عمیق، یادگیری ماشین،

فهرست مطالب

چکیده .....أ

فصل ۱ : روش پژوهش

۱-۱ مقدمه 2

۱-۲ پیشینه پژوهشی..... 2

فصل ۲ : مبانی نظری

۲-۱ مقدمه 5

۲-۲ مروری بر خلاصه‌سازی متن..... 5

۲-۳ انواع خلاصه‌سازی 6

۲-۳-۱ خلاصه سازی استخراجی 6

۲-۳-۲ خلاصه سازی انتزاعی 7

۲-۳-۳ مقایسه روش استخراجی و انتزاعی..... 8

۲-۴ پیشرفت های پردازش زبان طبیعی 8

۲-۴-۱ توسعه‌های اولیه درNLP 8

۲-۴-۲ ظهور یادگیری ماشین و یادگیری عمیق 9

۲-۴-۳ ترنسفورمرها و مدل‌های زبانی پیش‌آموخته 9

۲-۴-۴ نوآوری‌ها و کاربردهای اخیر 10

۲-۵ ترانسفورمر ها در خلاصه سازی متن 11

۲-۵-۱ معماری ترنسفورمرها 11

۲-۵-۲ نحوه عملکرد ترنسفورمرها 12

۲-۶ کارهای مرتبط در خلاصه‌سازی مقالات علمی 13

فصل ۳ : روش پژوهش

۳-۱ مقدمه 15
۳-۲ گردآوری داده ها 15
۳-۲-۱ توصیف مجموعه داده ها 15
۳-۲-۲ مراحل پیش‌پردازش 16
۳-۳ انتخاب مدل 17
۳-۳-۱ مروری بر مدل های ترنسفورمر 17
۳-۳-۲ مدل های انتخاب شده 17
۳-۴ جزئیات پیاده‌سازی 19
۳-۴-۱ کتابخانه های مورد استفاده 19
۳-۴-۲ انتخاب هایپرپارامتر های مدل 20
۳-۵ معیار های ارزیابی 20
۳-۵-۱ امتیاز ROUGE 20
فصل ۴ : نتیجه‌گیری و پیشنهاد ها
۴-۱ نتیجه گیری 23
۴-۲ پیشنهاد ها 24
مراجع 26

فصل ۱

معرفی پژوهش

۱-۱ مقدمه

در عصر حاضر، با گسترش روزافزون تولید اطلاعات علمی و افزایش حجم مقالات منتشر شده، نیاز به روش‌های مؤثر برای خلاصه‌سازی و استخراج اطلاعات کلیدی از این متون به شدت احساس می‌شود. خلاصه‌سازی متن به عنوان یک ابزار حیاتی برای تسهیل دسترسی به اطلاعات و ارتقاء کارایی پژوهشگران و دانشجویان در فرآیند مطالعه و تحلیل داده‌ها شناخته می‌شود. این تحقیق به بررسی روش‌های نوین خلاصه‌سازی متن، به ویژه در زمینه مقالات علمی، می‌پردازد و بر استفاده از مدل‌های پیشرفته ترانسفورمر تأکید دارد. این مدل‌ها به دلیل قابلیت‌های بالای خود در درک و تولید زبان طبیعی، پتانسیل بالایی برای بهبود کیفیت خلاصه‌ها دارند. در این راستا، هدف اصلی این پژوهش توسعه و ارزیابی الگوریتم‌های خلاصه‌سازی است که بتوانند به‌طور مؤثر و دقیق، اطلاعات کلیدی را از مقالات علمی استخراج کنند و در نهایت، به بهبود فرآیند یادگیری و پژوهش در این حوزه کمک نمایند.

۱-۲ پیشینه پژوهشی

خلاصه‌سازی مقالات علمی به عنوان یک چالش مهم در پردازش زبان طبیعی (NLP) شناخته می‌شود، که به دلیل حجم بالای اطلاعات و پیچیدگی‌های موجود در متون علمی، نیاز به روش‌های مؤثر و کارآمد برای استخراج اطلاعات کلیدی را افزایش می‌دهد. با توجه به پیشرفت‌های اخیر در یادگیری عمیق و به‌ویژه مدل‌های ترانسفورمر، این روش‌ها به عنوان ابزارهای جدید برای بهبود کیفیت خلاصه‌سازی متن معرفی شده‌اند.

مقاله "Abstractive text summarization: State of the art, challenges, and improvements" منتشر شده در ScienceDirect ، تکنیک‌های کلیدی در خلاصه‌سازی متن انتزاعی، مانند مدل‌های دنباله به دنباله را بررسی می‌کند، و چالش‌هایی مانند سازگاری واقعی را برجسته می‌کند و بر استراتژی‌های نوآورانه برای پیشبرد پیشرفت‌های آینده تأکید می‌کند.

مقاله " Fine Tuning Pretrained Transformers for Abstractive News Summarization " منتشر شده در IEEE ، مدل های T5، GPT-2، و BART را برای خلاصه‌سازی اخبار با استفاده از تکنیک‌هایی مانند یادگیری انتقال مقایسه می‌کند .همچنین این مقاله به چالش های بهینه سازی هایپرپارامترها اشاره می کند و نشان می‌دهد که مدل T5 در مجموعه داده‌های CNN/Daily Mail عملکرد بهتری نسبت به سایر مدل‌ها داشته است.

با این حال، علی‌رغم پیشرفت‌های اخیر، هنوز چالش‌هایی در این حوزه وجود دارد. بسیاری از روش‌های موجود به محدودیت‌های خاصی در زمینه دقت و قابلیت تعمیم برخورد می‌کنند و نیاز به داده‌های آموزشی بزرگ و متنوع دارند. همچنین، نواقص موجود در ارزیابی کیفیت خلاصه‌ها و عدم توجه به زمینه‌های خاص مقالات علمی، از دیگر مشکلاتی است که در تحقیقات پیشین دیده شده است. این تحقیق با هدف پر کردن این شکاف‌ها و بهبود فرآیند خلاصه‌سازی مقالات علمی از طریق فاین‌تیون کردن ترانسفورمرهای از پیش آموزش‌دیده‌شده، به بررسی و تحلیل روش‌های مختلف می‌پردازد و به دنبال ارائه راهکارهایی برای افزایش دقت و کیفیت خلاصه‌ها در این زمینه است.

فصل ۲

مبانی نظری

۲-۱ مقدمه

این فصل به بررسی و تحلیل پیشینه تحقیق در زمینه خلاصه‌سازی متن می‌پردازد. با توجه به افزایش روزافزون حجم اطلاعات علمی و نیاز به استخراج سریع و مؤثر اطلاعات کلیدی، خلاصه‌سازی متن به عنوان یک ابزار ضروری در پردازش زبان طبیعی شناخته می‌شود. این فصل به دو نوع اصلی خلاصه‌سازی، یعنی استخراجی و انتزاعی، می‌پردازد و به بررسی مزایا و معایب هر یک می‌پردازد. همچنین، تکنیک‌های نو و پیشرفته‌ای که در سال‌های اخیر در این زمینه توسعه یافته‌اند، از جمله مدل‌های ترانسفورمر و یادگیری عمیق، مورد بررسی قرار می‌گیرند. در این راستا، چالش‌ها و محدودیت‌های موجود در روش‌های خلاصه‌سازی نیز تحلیل می‌شود تا زمینه‌ای برای شناسایی نقاط ضعف و فرصت‌های بهبود فراهم گردد. هدف این فصل، ارائه یک نمای کلی از وضعیت کنونی تحقیقات در حوزه خلاصه‌سازی متن و شناسایی روندهای نوظهور و زمینه‌های تحقیقاتی آینده است که می‌تواند به پیشرفت‌های بیشتر در این حوزه کمک کند.

۲-۲ مروری بر خلاصه‌سازی متن

خلاصه‌سازی متن به عنوان یک تکنیک کلیدی در پردازش زبان طبیعی (NLP) شناخته می‌شود که هدف آن تولید نسخه‌ای مختصر از یک متن طولانی است، بدون از دست دادن اطلاعات اصلی و مهم آن. با توجه به رشد سریع حجم داده‌های متنی در دنیای دیجیتال، نیاز به روش‌های مؤثر برای خلاصه‌سازی اطلاعات به شدت احساس می‌شود .به طور کلی، خلاصه‌سازی متن به دو نوع اصلی تقسیم می‌شود: خلاصه‌سازی استخراجی و خلاصه‌سازی انتزاعی. در خلاصه‌سازی استخراجی، جملات یا عبارات مهم مستقیماً از متن اصلی انتخاب و در یک خلاصه ترکیب می‌شوند، در حالی که در خلاصه‌سازی انتزاعی ، ایده‌های کلیدی متن اصلی به صورت جملات جدیدی تولید می‌شوند که ممکن است عبارات و جملات جدیدی را شامل شوند .خلاصه‌سازی استخراجی به دلیل سادگی و کارایی خود، به طور گسترده‌ای مورد استفاده قرار می‌گیرد، اما محدودیت‌هایی نیز دارد، از جمله عدم توانایی در تولید جملات جدید و ناتوانی در درک عمیق‌تر متن .از سوی دیگر، خلاصه‌سازی انتزاعی به دلیل قابلیت تولید جملات جدید و انعطاف‌پذیری بیشتر در بیان مفاهیم، به طور فزاینده‌ای محبوب شده است و به عنوان یک حوزه تحقیقاتی فعال در نظر گرفته می‌شود .این فصل به بررسی روش‌ها و تکنیک‌های مختلف خلاصه‌سازی متن، چالش‌ها و فرصت‌های موجود در این حوزه می‌پردازد و سعی دارد تا زمینه‌ای برای تحقیقات آینده فراهم آورد. با توجه به چالش‌هایی مانند عدم دقت در نمایندگی معنایی و نیاز به ارزیابی‌های دقیق‌تر، این مطالعه به دنبال شناسایی روندهای نوظهور و راهکارهای بهبود در خلاصه‌سازی متن است.

۲-۳ انواع خلاصه‌سازی

خلاصه‌سازی متن یک حوزه مهم در پردازش زبان طبیعی (NLP) است که هدف آن فشرده‌سازی حجم‌های بزرگ متن به خلاصه‌های کوتاه و منسجم در حالی که اطلاعات اساسی حفظ می‌شود، است. دو نوع عمده تکنیک‌های خلاصه‌سازی شامل **خلاصه‌سازی استخراجی** و **خلاصه‌سازی انتزاعی** هستند. هر روش از الگوریتم‌ها و رویکردهای متفاوتی برای تولید خلاصه‌ها استفاده می‌کند و درک تفاوت‌های آن‌ها برای انتخاب تکنیک مناسب برای کاربردهای خاص بسیار مهم است.

۲-۳-۱ خلاصه سازی استخراجی

خلاصه‌سازی استخراجی شامل انتخاب و ترکیب جملات یا عبارات مرتبط‌ترین از متن اصلی برای ایجاد یک خلاصه است. این روش بر این فرض استوار است که جملات اطلاعاتی‌ترین در متن اصلی از قبل موجود هستند. الگوریتم‌های مورد استفاده برای خلاصه‌سازی استخراجی را می‌توان به چندین رویکرد تقسیم کرد :

- **روش‌های مبتنی بر فراوانی** :این روش‌ها، مانند فراوانی اصطلاحات-معکوس فراوانی سند(TF-IDF) ، اهمیت جملات را بر اساس فراوانی اصطلاحاتی که در آن‌ها وجود دارد، ارزیابی می‌کنند. جملات با فراوانی بالاتر از اصطلاحات به عنوان مهم‌تر در نظر گرفته می‌شوند و برای گنجاندن در خلاصه انتخاب می‌شوند .
- **روش‌های مبتنی بر گراف** :تکنیک‌هایی مانند TextRank و LexRank از ساختارهای گرافی برای نمایش جملات به عنوان گره‌ها و روابط آن‌ها به عنوان لبه‌ها استفاده می‌کنند. اهمیت هر جمله بر اساس ارتباط آن با سایر جملات در گراف تعیین می‌شود و جملات مرکزی‌تر برای خلاصه انتخاب می‌شوند. این روش‌ها از الگوریتم‌های PageRank که در موتورهای جستجوی وب استفاده می‌شوند، الهام گرفته‌اند .
- **رویکردهای یادگیری ماشین** :تکنیک‌های یادگیری نظارت‌شده نیز می‌توانند به کار گرفته شوند، جایی که مدل‌ها بر روی مجموعه‌های داده برچسب‌گذاری شده آموزش می‌بینند تا جملات کلیدی را شناسایی کنند. الگوریتم‌هایی مانند ماشین‌های بردار پشتیبان (SVM) و شبکه‌های عصبی برای طبقه‌بندی جملات بر اساس مرتبط بودن آن‌ها به کار رفته‌اند .

در حالی که خلاصه‌سازی استخراجی ساده است و معمولاً خروجی‌های گرامری درستی تولید می‌کند، محدودیت‌هایی نیز دارد. خلاصه‌های حاصل ممکن است فاقد انسجام و روانی باشند، زیرا آن‌ها صرفاً از متن اصلی بدون بازنویسی یا ترکیب به هم چسبیده‌اند. علاوه بر این، روش‌های استخراجی ممکن است اطلاعات مهمی که به‌طور صریح در جملات انتخاب‌شده بیان نشده‌اند را از دست بدهند.

۲-۳-۲ خلاصه سازی انتزاعی

در مقابل، خلاصه‌سازی انتزاعی جملات جدیدی تولید می‌کند که جوهره متن اصلی را در بر می‌گیرد. این رویکرد به دنبال درک محتوا و تولید خلاصه‌ای است که ممکن است شامل بازنویسی یا عبارات جدید باشد. خلاصه‌سازی انتزاعی شامل الگوریتم‌های پیچیده‌تری است، از جمله :

- **مدل‌های توالی به توالی** :این مدل‌ها که معمولاً بر اساس شبکه‌های عصبی بازگشتی (RNN) یا شبکه‌های حافظه بلندمدت و کوتاه‌مدت (LSTM) هستند، متن ورودی را به یک بردار با طول ثابت کدگذاری کرده و آن را به یک خلاصه رمزگشایی می‌کنند. مکانیزم توجه معمولاً برای تمرکز بر بخش‌های مرتبط ورودی در حین رمزگشایی به کار می‌رود .

- **ترنسفورمرها** :ظهور معماری‌های ترنسفورمر، مانند BERT وGPT ، پیشرفت‌های قابل توجهی را در خلاصه‌سازی انتزاعی به ارمغان آورده است. ترنسفورمرها از مکانیزم‌های خودتوجه برای به‌دست‌آوردن وابستگی‌های بلندمدت در متن استفاده می‌کنند و به آن‌ها امکان تولید خلاصه‌های منسجم‌تر و مرتبط‌تر با متن را می‌دهند .مدل‌هایی مانند T5 و BART نتایج چشمگیری در تولید خلاصه‌های شبیه به انسان نشان داده‌اند .

- **رویکردهای ترکیبی** :برخی از مطالعات اخیر تکنیک‌های استخراجی و انتزاعی را ترکیب می‌کنند تا از نقاط قوت هر دو روش بهره‌برداری کنند. به عنوان مثال، یک مرحله استخراجی ممکن است ابتدا جملات کلیدی را شناسایی کند و سپس یک مرحله انتزاعی این جملات را به یک خلاصه منسجم بازنویسی و ترکیب کند .

خلاصه‌سازی انتزاعی معمولاً خلاصه‌های روان‌تر و شبیه به انسان‌تری تولید می‌کند، اما پیاده‌سازی آن نیز چالش‌برانگیزتر است. مدل‌ها نیاز به آموزش گسترده بر روی مجموعه‌های داده بزرگ دارند و ممکن است با دقت واقعی مشکل داشته باشند و گاهی اطلاعاتی تولید کنند که در متن اصلی وجود ندارد.

۲-۳-۳ مقایسه روش استخراجی و انتزاعی

ویژگی	خلاصه سازی استخراجی	خلاصه سازی انتزاعی
روش	جملات را به‌طور مستقیم از متن انتخاب می‌کند.	جملات جدیدی بر اساس درک تولید می‌کند.
پیچیدگی	ساده‌تر، وابسته به انتخاب جملات	پیچیده‌تر، شامل تولید زبان است
انسجام	ممکن است به دلیل استخراج مستقیم فاقد انسجام باشد	معمولاً انسجام و روانی بیشتری دارد
دقت	بالا، زیرا از جملات اصلی استفاده می‌کند	متغیر، ممکن است نادرستی‌هایی را معرفی کند
منابع محاسباتی	به منابع بیشتری برای آموزش و استنباط نیاز دارد	معمولاً به منابع محاسباتی کمتری نیاز دارد

۲-۴ پیشرفت های پردازش زبان طبیعی

پردازش زبان طبیعی (NLP) از زمان تأسیس خود تحولات قابل توجهی را تجربه کرده است و از سیستم‌های ابتدایی مبتنی بر قوانین به مدل‌های پیچیده‌ای که قادر به درک و تولید متنی شبیه به انسان هستند، تکامل یافته است. این بخش تحولات کلیدی در NLP را بررسی می‌کند و پیشرفت‌های بنیادی و نوآوری‌های اخیر را که این حوزه را شکل داده‌اند، مورد توجه قرار می‌دهد.

۲-۴-۱ توسعه‌های اولیه درNLP

ریشه‌های NLP به دهه ۱۹۵۰ و ۱۹۶۰ برمی‌گردد، زمانی که تحقیقات اولیه بر روی ترجمه ماشینی و تجزیه نحوی متمرکز بود. یکی از سیستم‌های اولیه قابل توجه، آزمایش Georgetown-IBM در سال ۱۹۵۴ بود که توانایی ترجمه جملات روسی به انگلیسی را با استفاده از واژگان و قوانین گرامری محدود نشان داد. این دوره پایه‌گذار تحقیقات آینده در NLP بود، اما با چالش‌هایی به دلیل پیچیدگی زبان انسانی و محدودیت‌های قدرت محاسباتی مواجه بود .در دهه‌های ۱۹۷۰ و ۱۹۸۰، معرفی دستور زبان‌های رسمی، مانند دستور زبان تولیدی چامسکی، چارچوب نظری برای درک ساختار زبان فراهم کرد. با این حال، این رویکردهای مبتنی بر قوانین با ابهام و تنوع در زبان طبیعی دست و پنجه نرم کردند و منجر به بررسی روش‌های آماری در دهه ۱۹۹۰ شدند. ظهور NLP آماری یک تغییر پارادایمی را به همراه داشت، به طوری که پژوهشگران شروع به استفاده از مجموعه‌های بزرگ متنی برای توسعه مدل‌های احتمالی برای وظایفی مانند برچسب‌گذاری بخش‌های گفتار و شناسایی موجودیت‌های نام‌دار (NER) کردند.

۲-۴-۲ ظهور یادگیری ماشین و یادگیری عمیق

اوایل دهه ۲۰۰۰ شاهد ادغام تکنیک‌های یادگیری ماشین در NLP بود که به طور قابل توجهی عملکرد در زمینه‌های مختلف را بهبود بخشید. ماشین‌های بردار پشتیبان (SVM) و میدان‌های تصادفی شرطی (CRF) به عنوان ابزارهای محبوب برای وظایف طبقه‌بندی تبدیل شدند، در حالی که الگوریتم‌هایی مانند مدل‌های مخفی مارکوف (HMM) برای وظایف برچسب‌گذاری توالی به کار گرفته شدند .این روش‌ها به سیستم‌ها این امکان را می‌دادند که از داده‌ها یاد بگیرند و نه تنها به قوانین دست‌نویس وابسته باشند، که منجر به توسعه برنامه‌های NLP قوی‌تر و سازگارتر شد .ورود یادگیری عمیق در دهه ۲۰۱۰ انقلاب بزرگی در NLP ایجاد کرد .شبکه‌های عصبی، به ویژه شبکه‌های عصبی بازگشتی (RNN) و انواع آن‌ها، مانند شبکه‌های حافظه بلندمدت و کوتاه‌مدت(LSTM) ، امکان بهتری برای مدیریت داده‌های توالی‌دار فراهم کردند و آن‌ها را برای وظایفی مانند مدل‌سازی زبان و ترجمه ماشینی مناسب ساختند. کار برجسته باهدانائو و همکاران(۲۰۱۴) مکانیزم توجه را معرفی کرد که کیفیت ترجمه را به طور قابل توجهی بهبود بخشید و به مدل‌ها این امکان را داد که بر بخش‌های مرتبط از توالی ورودی تمرکز کنند .

۲-۴-۳ ترنسفورمرها و مدل‌های زبانی پیش‌آموخته

انتشار معماری ترنسفورمر توسط واسوانی و همکاران در سال ۲۰۱۷ نقطه عطفی در NLP بود. ترنسفورمرها از مکانیزم‌های خودتوجه برای پردازش داده‌های ورودی به‌صورت موازی استفاده می‌کنند که به طور قابل توجهی کارایی آموزش و عملکرد را در طیف وسیعی از وظایف بهبود می‌بخشد. این معماری زمینه‌ساز توسعه مدل‌های زبانی پیش‌آموخته شد که به سنگ بنای NLP مدرن تبدیل شده‌اند .مدل‌هایی مانند BERT (نمایش‌های دوجهته از ترنسفورمرها) و GPT (ترنسفورمر پیش‌آموخته تولیدی) قابلیت‌های قابل توجهی در درک زمینه و تولید متن منسجم نشان داده‌اندBERT ،. که توسط دوولین و همکاران (۲۰۱۸) معرفی شد، از تکنیک مدل‌سازی زبان ماسک‌شده استفاده می‌کند که به آن اجازه می‌دهد نمایش‌های دوجهته‌ای از متن یاد بگیرد و منجر به عملکرد پیشرفته در معیارهای مختلف شود . به‌طور مشابه، GPT-3که توسط OpenAI منتشر شد، پتانسیل مدل‌های زبانی در مقیاس بزرگ را نشان داد و قادر به تولید متنی شبیه به انسان در پاسخ به درخواست‌های متنوع و کاربردهای مختلف بود .

۲-۴-۴ نوآوری‌ها و کاربردهای اخیر

در سال‌های اخیر، حوزه NLP به تکامل خود ادامه داده و پیشرفت‌های قابل توجهی در چندین زمینه داشته است :

- **هوش مصنوعی مکالمه‌ای** :توسعه چت‌بات‌ها و دستیاران مجازی پیشرفته، تعاملات مشتری را در صنایع مختلف متحول کرده است. این سیستم‌ها از NLP برای درک پرسش‌های کاربر و ارائه پاسخ‌های مرتبط استفاده می‌کنند و تجربه کاربری را بهبود می‌بخشند .
- **تولید زبان طبیعی** NLG :**(NLG)** در ایجاد محتوای خودکار به محبوبیت رسیده و به سازمان‌ها این امکان را می‌دهد که متن‌های شبیه به انسان را از داده‌های ساختاری تولید کنند. این فناوری به‌ویژه در روزنامه‌نگاری و مالی، جایی که گزارش‌دهی به‌موقع و دقیق حیاتی است، مفید است .
- **تحلیل احساسات و شناسایی عواطف** :پیشرفت‌ها در تحلیل احساسات توانایی سیستم‌ها را در درک و پاسخ به تن‌های احساسی در ارتباطات بهبود بخشیده است. این قابلیت به‌طور فزاینده‌ای در خدمات مشتری و نظارت بر رسانه‌های اجتماعی مورد استفاده قرار می‌گیرد .
- **یادگیری چندمدالیته‌ای** :تحقیقات اخیر به بررسی ادغام NLP با سایر مدالیته‌ها، مانند داده‌های بصری و شنیداری پرداخته است که منجر به سیستم‌های درک جامع‌تری می‌شود. این روند با مدل‌هایی که می‌توانند هم متن و هم تصاویر را پردازش کنند تا خروجی‌های مرتبط با زمینه تولید کنند، به خوبی نمایان است .
- **ملاحظات اخلاقی و کاهش تعصب** :با گسترش فناوری‌هایNLP ، پرداختن به نگرانی‌های اخلاقی، از جمله تعصب در مدل‌های زبانی و استقرار مسئولانه سیستم‌هایNLP ، به موضوعی مهم تبدیل شده است. تحقیقات جاری به توسعه چارچوب‌هایی برای اطمینان از انصاف و پاسخگویی در کاربردهای NLP می‌پردازد .

پیشرفت‌ها در پردازش زبان طبیعی به‌طور قابل توجهی شیوه تعامل ماشین‌ها با زبان انسانی را متحول کرده است. از سیستم‌های ابتدایی مبتنی بر قوانین تا مدل‌های پیشرفته یادگیری عمیق کنونی، این حوزه شاهد پیشرفت‌های قابل توجهی بوده است. با ادامه تکاملNLP ، تحقیقات و نوآوری‌های مداوم برای مقابله با چالش‌ها و باز کردن امکانات جدید برای کاربردها در زمینه‌های مختلف ضروری خواهد بود.

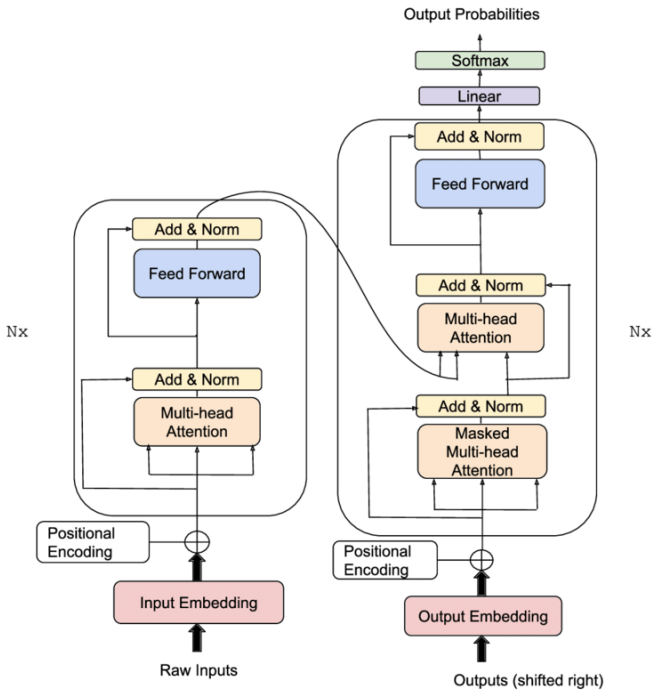
۲-۵ ترانسفورمر ها در خلاصه سازی متن

ترنسفورمرها انقلاب بزرگی در حوزه پردازش زبان طبیعی (NLP) به‌ویژه در زمینه خلاصه‌سازی متن ایجاد کرده‌اند. معماری منحصر به فرد و مکانیزم توجه آن‌ها به آن‌ها این امکان را می‌دهد که روابط بین کلمات در یک توالی را به‌طور مؤثری شناسایی کنند، که این امر آن‌ها را نسبت به مدل‌های قبلی مانند شبکه‌های عصبی بازگشتی (RNN)و شبکه‌های حافظه بلندمدت و کوتاه‌مدت (LSTM) برتر می‌سازد. این بخش به بررسی معماری ترنسفورمرها، نحوه عملکرد آن‌ها و کاربردشان در خلاصه‌سازی متن می‌پردازد.

۲-۵-۱ معماری ترنسفورمرها

معماری ترنسفورمر که در مقاله "توجه، همه آن چیزی است که نیاز دارید" توسط واسوانی و همکاران (2017) معرفی شد، شامل یک ساختار کدگذار-رمزگذار است.

کدگذار متن ورودی را پردازش کرده و آن را به یک نمایش پیوسته تبدیل می‌کند که اطلاعات اساسی را در بر می‌گیرد. این بخش شامل چندین لایه است که هر یک از دو مؤلفه اصلی تشکیل شده‌اند: توجه خودی چندسر و یک شبکه عصبی پیش‌خور. رمزگذار توالی خروجی (خلاصه) را از نمایش کدگذاری شده تولید می‌کند. این بخش نیز شامل چندین لایه است و دارای یک مکانیزم توجه اضافی است که به آن اجازه می‌دهد بر بخش‌های مرتبط با ورودی تمرکز کند.



۲-۵-۲ نحوه عملکرد ترنسفورمرها

نوآوری اصلی ترنسفورمرها در مکانیزم توجه آن‌ها نهفته است که به مدل این امکان را می‌دهد تا اهمیت کلمات مختلف در یک توالی را هنگام تولید یک خلاصه ارزیابی کند. این کار از طریق مراحل زیر انجام می‌شود:

۱. **تبدیل ورودی به بردار** :متن ورودی ابتدا به نمایش‌های برداری تبدیل می‌شود که از یک لایه تعبیه استفاده می‌کند. هر کلمه به یک بردار پیوسته که معنی معنایی آن را در بر می‌گیرد، نگاشته می‌شود.

۲.**کدگذاری موقعیتی** :از آنجا که ترنسفورمرها به‌طور ذاتی ترتیب کلمات را درک نمی‌کنند، کدگذاری موقعیتی به تعبیه‌ها اضافه می‌شود. این کدگذاری از توابع سینوس و کسینوس برای ارائه اطلاعات در مورد موقعیت هر کلمه در توالی استفاده می‌کند.

۳. **مکانیزم توجه خودی** :مکانیزم توجه خودی به مدل این امکان را می‌دهد که روابط بین کلمات در توالی ورودی را محاسبه کند. برای هر کلمه، سه بردار ایجاد می‌شود: پرسش(QQ) ، کلید (KK) و ارزش .(VV) نمرات توجه با استفاده از حاصل‌ضرب نقطه‌ای بین بردارهای پرسش و کلید محاسبه می‌شود:

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

در اینجا، d_k بعد بردارهای کلید است و تابع softmax برای به‌دست‌آوردن وزن‌های توجه که تعیین می‌کند هر کلمه چقدر باید بر اساس ارتباطش با کلمات دیگر تمرکز کند، اعمال می‌شود.

۴. **توجه چندسر** :به‌جای انجام یک تابع توجه، ترنسفورمرها از چندین سر توجه برای شناسایی انواع مختلف روابط استفاده می‌کنند. هر سر به‌طور مستقل ورودی را پردازش می‌کند و به مدل این امکان را می‌دهد که جنبه‌های مختلف داده را یاد بگیرد. خروجی‌های همه سرها ترکیب و به‌صورت خطی تبدیل می‌شوند تا خروجی نهایی لایه توجه را تشکیل دهند.

۵. **شبکه پیش‌خور** :پس از لایه توجه، خروجی از طریق یک شبکه عصبی پیش‌خور عبور می‌کند که شامل دو تبدیل خطی با یک تابع فعال‌سازی ReLU در بین آن‌ها است. این شبکه اطلاعات را بیشتر پردازش کرده و توانایی مدل را برای شناسایی الگوهای پیچیده افزایش می‌دهد.

۶. **اتصالات باقی‌مانده و نرمال‌سازی لایه** :برای تسهیل آموزش، اتصالات باقی‌مانده به دور هر زیرلایه (توجه و پیش‌خور) اضافه می‌شود و نرمال‌سازی لایه برای تثبیت فرآیند یادگیری اعمال می‌شود.

در خلاصه‌سازی متن، ترنسفورمرها به‌دلیل توانایی‌شان در در نظر گرفتن کل زمینه متن ورودی، عملکرد فوق‌العاده‌ای دارند. کدگذار کل توالی ورودی را پردازش می‌کند و روابط بین تمام کلمات را شناسایی می‌کند، در حالی که رمزگذار خلاصه را با تمرکز بر بخش‌های مرتبط با ورودی تولید می‌کند .ترنسفورمرها می‌توانند بر روی مجموعه‌های بزرگ داده‌ای از مقالات علمی یا مقالات آموزشی تنظیم شوند تا خلاصه‌های با کیفیت بالا تولید کنند.

۶-۲ کارهای مرتبط در خلاصه‌سازی مقالات علمی

در سال‌های اخیر، حوزه خلاصه‌سازی متن شاهد پیشرفت‌های قابل توجهی بوده است، به‌ویژه با معرفی مدل‌های ترنسفورمر. این مدل‌ها شیوه تولید خلاصه‌ها را متحول کرده و با استفاده از معماری منحصر به فرد و مکانیزم‌های توجه خود، روابط پیچیده درون متن را شناسایی می‌کنند. تحقیقات نشان داده‌اند که ترنسفورمرها، مانند BERT وGPT ، در وظایف خلاصه‌سازی استخراجی و انتزاعی عملکرد بسیار خوبی دارند و در مقایسه با روش‌های سنتی از نظر انسجام و مرتبط بودن برتری دارند. مطالعات مختلف به بررسی کاربرد ترنسفورمرها در خلاصه‌سازی مقالات علمی، مقالات خبری و سایر اشکال متن پرداخته‌اند و اثربخشی و تنوع آن‌ها را نشان داده‌اند. بررسی‌های مداوم در مورد رویکردهای مبتنی بر ترنسفورمر همچنان تکنیک‌ها و روش‌های جدیدی را آشکار می‌کند و به تحول در حوزه خلاصه‌سازی متن کمک می‌کند و معیارهای جدیدی برای تحقیقات آینده در این زمینه تعیین می‌کند.

لیو و لاپاتا روشی را برای خلاصه سازی متن با استفاده از مدل های BERT از پیش آموزش دیده پیشنهاد کردند. آنها یک رمزگذار جدید در سطح سند مبتنی بر BERT برای درک معنایی یک سند و به دست آوردن بازنمایی جملات معرفی کردند. رویکرد آنها شامل هر دو مدل خلاصه سازی استخراجی و انتزاعی است. مدل استخراجی از لایه های ترانسفورماتور بین جمله ای استفاده می کند، در حالی که مدل انتزاعی از یک برنامه تنظیم دقیق جدید با بهینه سازهای جداگانه برای رمزگذار و رمزگشا استفاده می کند. این روش بر روی سه مجموعه داده مورد ارزیابی قرار گرفت و به نتایج پیشرفته‌ای در هر دو محیط استخراجی و انتزاعی دست یافت.

خاندلوال و همکاران روشی را برای خلاصه سازی متن کارآمد با استفاده از یک ترانسفورماتور از پیش آموزش دیده پیشنهاد کرد. این رویکرد از یک شبکه فقط رمزگشا استفاده می کند که در آن همان Transformer LM منبع را رمزگذاری می کند و خلاصه را تولید می کند. با استفاده از نمرات ROUGE در مجموعه داده های CNN/Daily Mail مورد ارزیابی قرار گرفت که نشان دهنده پیشرفت های قابل توجهی در کارایی نمونه است. مدل ROUGE-2 13.1 را تنها با استفاده از 1 درصد از داده های آموزشی به دست آورد، که به طور قابل توجهی بهتر از مدل های رمزگذار-رمزگشا از قبل آموزش دیده عمل کرد.

فصل ۳

روش پژوهش

۳-۱ مقدمه

این پروژه به‌طور خاص بر روی استخراج و خلاصه‌سازی محتوا از فایل‌های XML تمرکز دارد و تحلیل مقایسه‌ای از مدل‌های مختلف خلاصه‌سازی، از جمله Longformer، BigBird-Pegasusو یک مدل فاین تیون شده را انجام می‌دهد.

این روند با با آماده‌سازی داده‌ها آغاز می‌شود، جایی که فایل‌های XML پردازش می‌شوند تا چکیده‌ها و محتوای اصلی مقالات استخراج شوند و سپس داده‌ها پاک‌سازی و به مجموعه‌های آموزشی، اعتبارسنجی و آزمایشی تقسیم می‌شوند. این کار اطمینان حاصل می‌کند که داده‌ها برای آموزش مؤثر مدل‌ها آماده هستند. فرآیند خلاصه‌سازی از کتابخانه ترنسفورمرهای Hugging Face استفاده می‌کند و مدل‌هایی را به کار می‌گیرد که برای مستندات طولانی و داده‌های متنی بزرگ بهینه‌سازی شده‌اند. عملکرد هر مدل از طریق شباهت کازین ارزیابی می‌شود که با استفاده از بردارهای TF-IDF محاسبه می‌شود تا ارزیابی کند که چقدر خلاصه‌های تولیدشده با متن اصلی هم‌راستا هستند.

علاوه بر این، تنظیم هایپرپارامترها برای بهینه‌سازی نرخ‌های یادگیری، اندازه‌های بچ و پیکربندی‌های مدل انجام می‌شود تا دقت خلاصه‌سازی به حداکثر برسد. هدف نهایی این است که این مدل‌ها را بهبود بخشیده و قابلیت‌های خلاصه‌سازی آن‌ها را افزایش دهیم، به‌طوری که استفاده از اطلاعات در زمینه‌های مختلف، از جمله تحقیقات علمی، تسهیل شود .

۳-۲ گردآوری داده ها

در مرحله گردآوری داده‌ها، فایل‌های XML پردازش شدند تا چکیده‌ها و محتوای اصلی مرتبط استخراج شود. سپس داده‌ها پاکسازی و به مجموعه‌های آموزشی، اعتبارسنجی و آزمایشی تقسیم شدند تا یک مجموعه داده قوی برای آموزش و ارزیابی مدل‌های خلاصه‌سازی فراهم شود.

۳-۲-۱ توصیف مجموعه داده ها

مجموعه داده‌ای که در این پروژه استفاده شده ، از PubMed Central® (PMC) استخراج شده است، که یک مخزن جامع از ژورنال های علمی مجلات بیومدیکال است که توسط کتابخانه ملی پزشکی مؤسسه‌های ملی بهداشت ایالات متحده نگهداری می‌شود. PMC به عنوان یک منبع حیاتی برای تحقیقات معاصر بیومدیکال و مراقبت‌های بهداشتی عمل می‌کند و دسترسی به مجموعه وسیعی از مقالات را فراهم می‌آورد که از مطالعات جاری و تلاش‌های علمی آینده پشتیبانی می‌کند. این مجموعه داده شامل دامنه‌ای متنوع از مقالات در

فرمتXML است، که فرآیند استخراج و پردازش داده‌ها را برای خلاصه‌سازی تسهیل می‌کند. این فرمت ساختاریافته امکان تحلیل سیستماتیک چکیده‌ها و محتوای اصلی را فراهم می‌آورد و اطمینان حاصل می‌کند که مدل‌های خلاصه‌سازی می‌توانند به‌طور مؤثری بر روی داده‌های با کیفیت بالا آموزش ببینند.

۲-۲-۳ مراحل پیش‌پردازش

مرحله پیش‌پردازش این پروژه شامل مجموعه‌ای از مراحل برای آماده‌سازی مجموعه داده‌ها برای تنظیم و ارزیابی مدل بود.

اولین مرحله شامل دانلود انبوه فایل‌های XML مرتبط از PMC بود. این فایل‌ها سپس پردازش و به فرمت JSON تبدیل شدند که بر روی استخراج دو مؤلفه کلیدی تمرکز داشت: چکیده مقاله، که از بخش چکیده استخراج شده، و متن اصلی مقاله، که شامل متن کامل به‌جز چکیده است. این فرآیند تبدیل به‌منظور ساده‌سازی مراحل بعدی پاک‌سازی و آماده‌سازی داده‌ها انجام شد.

برای اطمینان از سازگاری، یکپارچگی و خوانایی، چندین تکنیک رایج پاک‌سازی داده‌ها به کار گرفته شد. کاراکترهای غیر ASCII از متن حذف شدند، برچسب‌های XML برای جداسازی محتوای مرتبط حذف شدند و فضاهای خالی و خط‌های جدید مناسب برای بهبود فرمت‌بندی اضافه شدند.

پس از پاک‌سازی اولیه، پنج رویکرد مختلف برای آماده‌سازی بیشتر داده‌ها برای تنظیم مدل به کار گرفته شد. این رویکردها در کاربرد لِماتیزاسیون، حذف کلمات توقف و نحوه برخورد با محتوای چکیده در متن اصلی متفاوت بودند. دلیل این رویکردها بررسی تأثیر تکنیک‌های مختلف پاک‌سازی بر عملکرد مدل و شناسایی بهترین ترکیب برای خلاصه‌سازی مقالات بیومدیکال بود.

رویکرد اول شامل لِماتیزاسیون هر دو متن چکیده و متن اصلی، حذف کلمات توقف از هر دو بخش و حذف محتوای چکیده از متن اصلی برای جلوگیری از تکرار بود. رویکرد دوم تنها لِماتیزاسیون و حذف کلمات توقف را بر روی متن اصلی اعمال کرد و همچنان چکیده را حذف کرد. رویکرد سوم متن کامل اصلی را شامل چکیده نگه داشت، اما لِماتیزاسیون و حذف کلمات توقف را بر روی هر دو بخش اعمال کرد. رویکرد چهارم تنها لِماتیزاسیون و حذف کلمات توقف را بر روی متن اصلی انجام داد و چکیده را حفظ کرد. در نهایت، رویکرد پنجم به سادگی محتوای چکیده را از متن اصلی حذف کرد.

با به‌کارگیری این رویکردهای مختلف پاک‌سازی، پروژه به بررسی تأثیر آن‌ها بر عملکرد مدل و تعیین بهترین ترکیب برای خلاصه‌سازی مقالات بیومدیکال پرداخت. مجموعه داده‌های پاک‌سازی‌شده که از این مراحل

پیش‌پردازش به‌دست آمدند، سپس برای تنظیم و ارزیابی عملکرد مدل‌های مختلف خلاصه‌سازی متن، از جمله Longformer، BigBird-Pegasusو مدل فاین تیون شده استفاده شد.

۳-۳ انتخاب مدل

در انتخاب مؤثرترین مدل برای خلاصه‌سازی مقالات بیومدیکال، چندین مدل پیش‌آموخته پیشرفته ارزیابی شدند. اگرچه BART به خاطر عملکرد قوی‌اش در تولید متن شناخته شده است، اما به دلیل محدودیت اندازه پنجره‌اش و حجم بالای داده‌های ورودی، برای مورد ما مناسب تشخیص داده نشد. در عوض، مدل‌هایی مانند-BigBird PEGASUS و Longformer به خاطر توانایی‌شان در مدیریت کارآمد مستندات طولانی ترجیح داده شدند.

۳-۳-۱ مروری بر مدل های ترنسفورمر

مدل‌های ترنسفورمر انقلاب بزرگی در زمینه پردازش زبان طبیعی (NLP) ایجاد کرده‌اند و امکان مدیریت کارآمد وظایف پیچیده‌ای مانند خلاصه‌سازی متن، ترجمه و تحلیل احساسات را فراهم می‌آورند. این مدل‌ها از مکانیزم توجه خودکار استفاده می‌کنند که به آن‌ها اجازه می‌دهد اهمیت کلمات مختلف در یک جمله را بدون توجه به موقعیت آن‌ها وزن‌دهی کنند. این قابلیت به ترنسفورمرها کمک می‌کند تا وابستگی‌های بلندمدت در متن را شناسایی کنند و آن‌ها را برای درک زمینه و تولید خروجی‌های منسجم به‌ویژه مؤثر می‌سازد .مدل‌های معروف ترنسفورمر شامل BERT است که در درک زمینه عالی است و GPT که برای تولید متن طراحی شده است. علاوه بر این، مدل‌هایی مانند BART و PEGASUSبه‌طور خاص برای وظایف خلاصه‌سازی بهینه‌سازی شده‌اند و عملکردی پیشرفته در بنچمارک‌های مختلف ارائه می‌دهند. انعطاف‌پذیری و قدرت معماری‌های ترنسفورمر همچنان به پیشرفت‌ها در NLP کمک می‌کند و برنامه‌های پیچیده‌تری را تسهیل می‌نماید و درک کلی زبان را بهبود می‌بخشد.

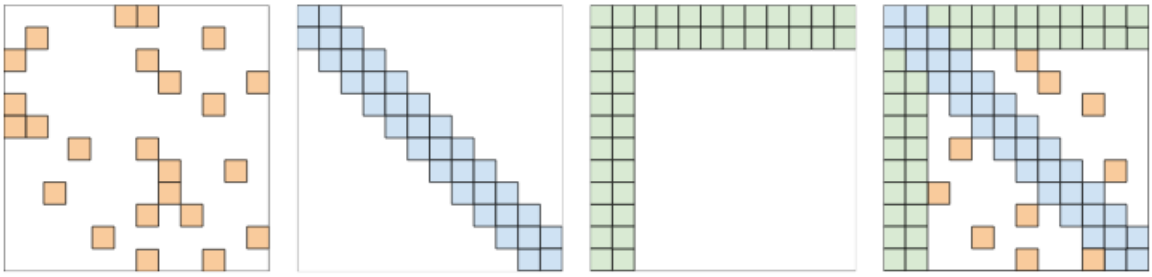
۳-۳-۲ مدل های انتخاب شده

مدل‌های منتخب برای خلاصه‌سازی شامل BigBird-PEGASUS است که در مدیریت مستندات طولانی خوب عمل می‌کند و Longformer که برای پردازش کارآمد متون بیومدیکال گسترده به منظور تولید خلاصه‌های اطلاعاتی بهینه‌سازی شده است.

الف) BigBird-PEGASUS

مدل BigBird-PEGASUS یک ترنسفورمر پیشرفته است که برای مدیریت توالی‌های طولانی متن طراحی شده است و آن را برای تسک های خلاصه‌سازی بسیار مؤثر می‌سازد. این مدل قدرت‌های BigBird که از مکانیزم توجه پراکنده استفاده می‌کند و PEGASUS که برای خلاصه‌سازی انتزاعی بهینه‌سازی شده است را ترکیب می‌کند.

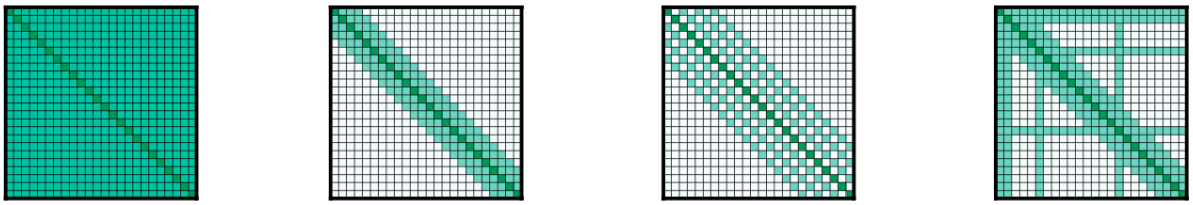
این مدل می‌تواند توالی‌هایی تا 4096 توکن را مدیریت کند که به طور قابل توجهی عملکرد را در تسک هایی که مستندات طولانی درگیر هستند، بهبود می‌بخشد. با کاهش پیچیدگی درجه دوم مکانیزم‌های توجه سنتی به خطی، BigBird-PEGASUS مقالات گسترده بیومدیکال را به طور کارآمد پردازش می‌کند و امکان تولید خلاصه‌های اطلاعاتی را فراهم می‌آورد. معماری آن امکان توجه جهانی و تصادفی را در کنار توجه پراکنده بلوکی فراهم می‌کند که توانایی آن را در شناسایی اطلاعات زمینه‌ای در سراسر متون بزرگ افزایش می‌دهد. این موضوع BigBird-PEGASUSرا به ابزاری قدرتمند برای پژوهشگران و متخصصان در حوزه پردازش زبان طبیعی تبدیل می‌کند.



(a) Random attention (b) Window attention (c) Global Attention (d) BIGBIRD

ب) Longformer

مدل Longformer یک معماری ترنسفورمر است که به‌طور خاص برای پردازش کارآمد مستندات طولانی طراحی شده است. این مدل به محدودیت‌های ترنسفورمرهای سنتی که در پردازش توالی‌های طولانی با مشکل مواجه می‌شوند، به‌ویژه به دلیل مکانیزم توجه خود که به‌صورت درجه دوم با طول توالی مقیاس می‌یابد، پاسخ می‌دهد . Longformerیک مکانیزم توجه نوآورانه معرفی می‌کند که توجه محلی (پنجره متحرک) و توجه جهانی را ترکیب می‌کند و به آن اجازه می‌دهد تا توالی‌هایی با هزاران توکن را بدون بار اضافی قابل توجه در حافظه و محاسبات پردازش کند. این مقیاس‌پذیری خطی Longformer را برای تسک هایی مانند خلاصه‌سازی متن و سوال و پاسخ مؤثر می‌سازد. این مدل در بنچمارک‌های مختلف عملکردی پیشرفته را نشان داده و مدل‌های قبلی مانند RoBERTa را در وظایف مرتبط با مستندات طولانی پشت سر گذاشته است و آن را به ابزاری ارزشمند در کاربردهای پردازش زبان طبیعی تبدیل می‌کند.



(a) Full n^2 attention (b) Sliding window attention (c) Dilated sliding window (d) Global+sliding window

۳-۴ جزئیات پیاده‌سازی

در این بخش به بررسی کتابخانه های مهم مورد استفاده در پروژه و هایپرپارامترهای انتخابی می‌پردازیم.

۳-۴-۱ کتابخانه های مورد استفاده

این پروژه از مجموعه‌ای از کتابخانه‌ها برای تسهیل پردازش داده، آموزش مدل و ارزیابی استفاده کرد. کتابخانه‌های NumPyو Pandasبرای محاسبات عددی کارآمد و دستکاری داده‌ها به‌کار گرفته شدند. NumPy برای عملیات آرایه و توابع ریاضی استفاده شد، Pandas امکان مدیریت داده‌های ساختاریافته در قالب DataFrame را فراهم کرد.

کتابخانه Transformers از Hugging Face نقش مهمی در این پروژه ایفا کرد. این کتابخانه دسترسی به مدل‌های پیش‌آموخته پیشرفته‌ای مانندBART ، BigBird-PEGASUSو Longformer را فراهم کرد که برای تسک خلاصه‌سازی متن بهینه‌سازی شدند. علاوه بر این، کلاس‌هایTokenizerاز Transformers برای پیش‌پردازش داده‌های متنی ورودی استفاده شدند تا سازگاری با مدل‌ها تضمین شود.

فریم‌ورک یادگیری عمیق PyTorch برای آموزش و استنتاج مدل‌ها به‌کار گرفته شد. این فریم‌ورک انعطاف‌پذیری در تعریف و آموزش مدل‌های کاستوم شده و همچنین استفاده از شتاب‌دهی GPU برای محاسبات کارآمد را فراهم کرد .کتابخانه Datasets از Hugging Face برای بارگذاری و مدیریت مجموعه داده استفاده شد که امکان بارگذاری و دسته‌بندی کارآمد داده‌ها در طول آموزش را فراهم کرد.

کتابخانه Natural Language Toolkit (NLTK) برای انجام تسک های مختلف پردازش متن، مانند توکن‌سازی، حذف کلمات توقف و لِماتیزاسیون استفاده شد .این تکنیک‌ها در طول پیش‌پردازش داده‌ها به‌منظور پاک‌سازی و آماده‌سازی متن برای آموزش مدل به‌کار گرفته شدند.

کتابخانه scikit-learn برای تقسیم داده‌ها استفاده شد که امکان تقسیم مجموعه داده به مجموعه‌های آموزشی، اعتبارسنجی و آزمایشی را فراهم کرد .این کار اطمینان حاصل کرد که مدل‌ها بر روی داده‌های نادیده ارزیابی شوند و ارزیابی قابل اعتمادی از عملکرد آن‌ها ارائه شود.

کتابخانه‌های دیگر مانند tqdm برای ردیابی پیشرفت، Richبرای فرمت‌بندی کنسول و Matplotlib برای رسم داده‌ها برای بهبود عملکرد و تجربه کاربری پروژه به‌کار گرفته شدند.

۳-۴-۲ انتخاب هایپرپارامتر های مدل

در این پروژه، تنظیم هایپرپارامترها مرحله مهمی در بهینه‌سازی عملکرد مدل‌های خلاصه‌سازی بود. کلاس Seq2SeqTrainingArguments 0020 از کتابخانه Transformers برای پیکربندی پارامترهای مختلف آموزش استفاده شد. دایرکتوری خروجی به "led-arxiv-10000" تنظیم شد که امکان ذخیره‌سازی منظم نتایج و چک‌پوینت‌های مدل را فراهم می‌کرد.

برای اطمینان از ارزیابی مؤثر، مدل به‌گونه‌ای پیکربندی شد که هر ۳۰۰ گام در طول آموزش ارزیابی شود، با استراتژی ارزیابی تنظیم شده بر روی "گام‌ها". این رویکرد بازخورد به موقع در مورد عملکرد مدل را تسهیل کرد . فرآیند آموزش طی پنج اپوک انجام شد، با اندازه دسته آموزشی ۱ و اندازه دسته ارزیابی ۱، که استفاده از حافظه و کارایی محاسباتی را متعادل می‌کرد.

نرخ یادگیری ۵e-۶ برای بهینه‌سازی مدل انتخاب شد. علاوه بر این، استفاده از اعداد اعشاری ۱۶ بیتی (fp16) به کاهش مصرف حافظه کمک کرد که امکان اندازه‌های دسته بزرگتر و افزایش سرعت آموزش را فراهم می‌کرد. تجمع گرادیان به ۴ گام تنظیم شد تا استفاده از حافظه را بهینه‌تر کند.

مدل به‌گونه‌ای طراحی شد که بهترین نسخه را هر ۹۰۰ گام ذخیره کند، با محدودیت کل ۵۰ مدل ذخیره‌شده. در نهایت، پیکربندی اطمینان حاصل کرد که بهترین مدل در پایان آموزش بارگذاری شود تا پتانسیل نتایج خلاصه‌سازی با کیفیت بالا را حداکثر کند.

۳-۵ معیار های ارزیابی

ارزیابی یک مدل خلاصه سازی شامل ارزیابی هر دو مدل از پیش آموزش دیده و مدل دقیق تنظیم شده است تا اطمینان حاصل شود که خلاصه های تولید شده دقیق، منسجم و مرتبط هستند.

۳-۵-۱ امتیاز ROUGE

امتیاز ROUGE (Recall-Oriented Understudy for Gisting Evaluation) مجموعه‌ای از معیارهای مهم است که برای ارزیابی کیفیت سیستم‌های خلاصه‌سازی خودکار و ترجمه ماشینی استفاده می‌شود. این امتیاز

همپوشانی بین یک خلاصه تولید شده توسط سیستم و یک یا چند خلاصه مرجع انسانی را اندازه‌گیری می‌کند. امتیاز ROUGE از 0 تا 1 متغیر است، به‌طوری که امتیاز نزدیک‌تر به 1 نشان‌دهنده شباهت بالاتر بین خلاصه تولید شده و خلاصه‌های مرجع است.

ROUGE شامل چندین نوع است، از جمله ROUGE-N که همپوشانیn -گرام‌ها (مانند یک‌گرام‌ها و دوگرام‌ها) را ارزیابی می‌کند و ROUGE-L که طولانی‌ترین زنجیره مشترک بین دو متن را بررسی می‌کند. این معیارها بینش‌هایی درباره اینکه یک مدل خلاصه‌سازی چقدر اطلاعات کلیدی را از متن اصلی حفظ می‌کند و در عین حال خروجی‌های مختصر تولید می‌کند، ارائه می‌دهند. امتیازهای بالای ROUGE معمولاً عملکرد بهتری را نشان می‌دهند و به همین دلیل برای مقایسه مدل‌های مختلف و بهبود قابلیت‌های خلاصه‌سازی در وظایف پردازش زبان طبیعی ضروری هستند.

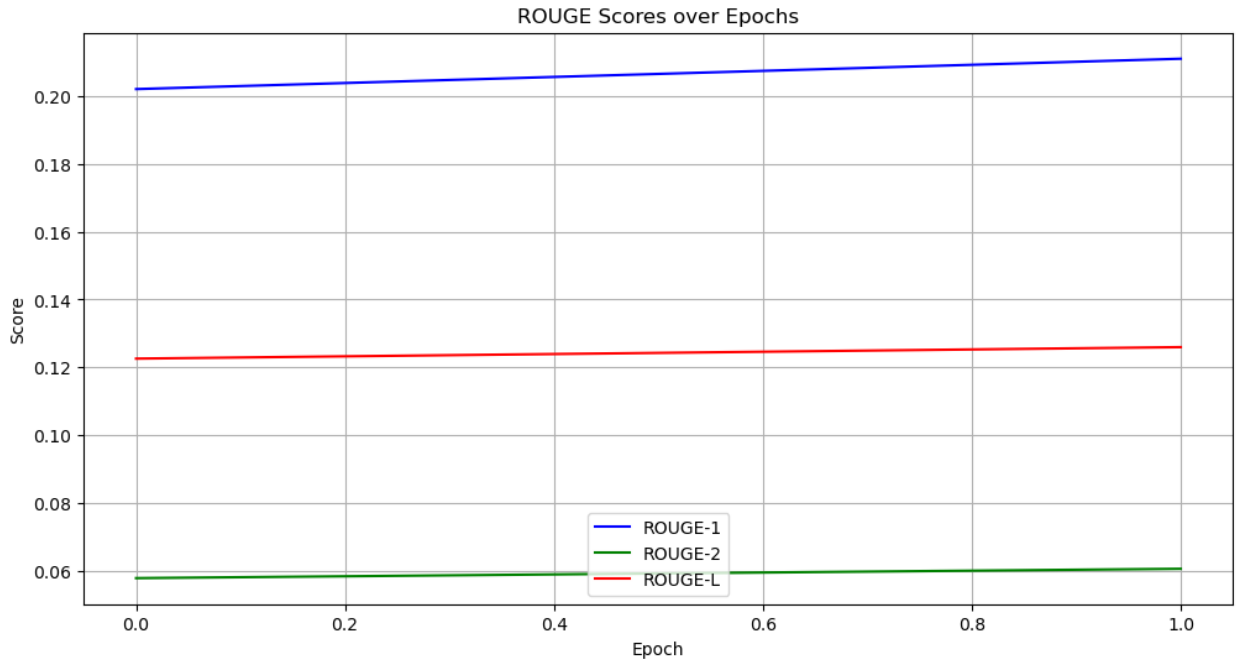
فصل ۴

نتیجه‌گیری و پیشنهاد ها

۴-۱ نتیجه گیری

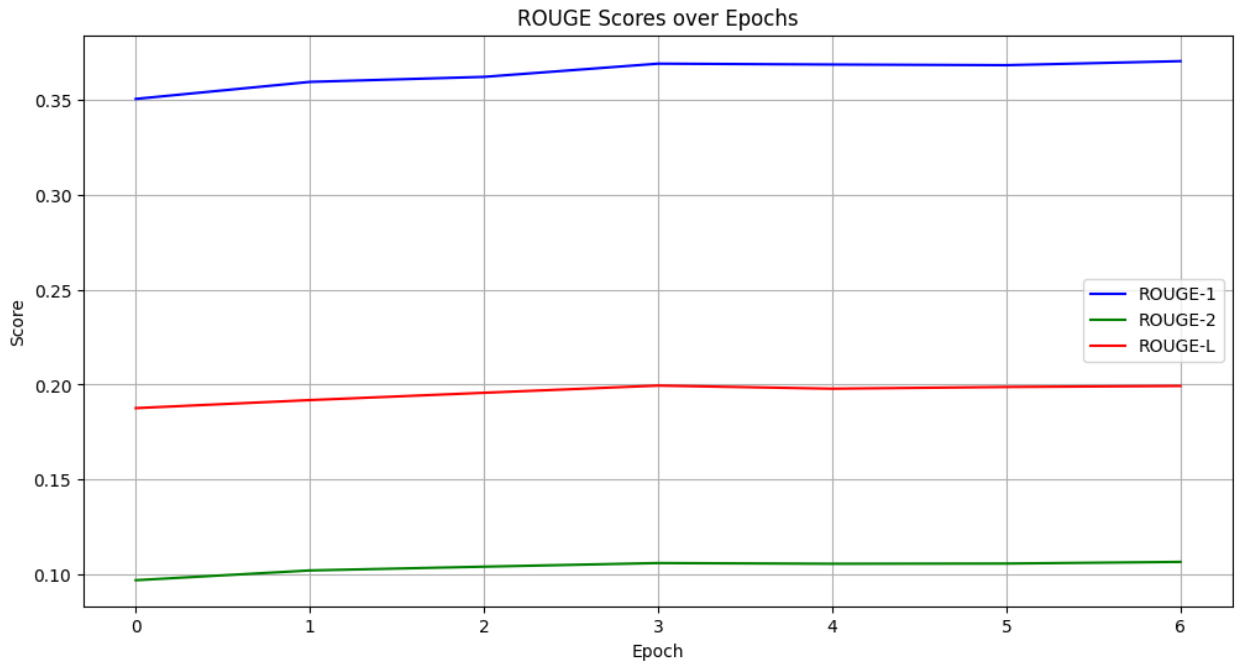
این پروژه یک مطالعه مقایسه‌ای از ترنسفورمرهای محبوب پردازش زبان طبیعی برای خلاصه‌سازی متون طولانی را ارائه داده است. آزمایش‌های انجام‌شده در این مطالعه شامل تنظیم هایپرپارامترها و انواع مختلف پیش‌پردازش داده ها بود که نتایج امیدوارکننده‌ای در زمینه خلاصه‌سازی متون طولانی داشت. با استفاده از امتیازROUGE ، عملکرد مدل‌های مختلف را ارزیابی کردیم و دریافتیم که حذف محتوای چکیده از بدنه متن در طی پیش‌پردازش، نتایج را به‌طور قابل‌توجهی بهبود بخشید. در کارهای آینده، بررسی رویکردهای ترکیبی که هر دو روش استخراجی و انتزاعی را ادغام می‌کنند، می‌تواند مزایای هر تکنیک را به‌کار گیرد و منجر به بهبود کیفیت خلاصه شود.

در دو تصویر زیر، نمودار های مربوط به امتیاز ROUGE را مشاهده می‌کنید که هر کدام مربوط به یکی از و نوع روش پیش‌پردازش داده ها هستند.



این نمودار نتیجه امتیاز ROUGE برای داده هایی است که با رویکرد اول پیش‌پردازش شده‌اند. رویکرد اول شامل لِماتیزاسیون هر دو متن چکیده و متن اصلی، حذف کلمات توقف از هر دو بخش و حذف محتوای چکیده از متن اصلی برای جلوگیری از تکرار بود.

نمودار زیر نیز نتیجه امتیاز ROUGE برای داده هایی است که با رویکرد پنجم پیش‌پردازش شده‌اند. رویکرد پنجم نیز به این صورت بود که محتوای چکیده را از متن اصلی حذف می‌کرد.



همانطور که مشاهده می‌کنید، شیب نمودار های مربوط به ROUGE-1 و ROUGE-2 و ROUGE-L در شکل دوم بیشتر است یعنی به صورت صعودی به عدد یک نزدیک تر میشود که این نشان‌دهنده‌ی شباهت بیشتر خلاصه تولید شده توسط مدل فاین تیون شده و خلاصه‌ی اصلی است.

۴-۲ پیشنهاد ها

در حالی که نتایج این پروژه اثربخشی حذف چکیده‌ها در طی پیش‌پردازش برای خلاصه‌سازی متون علمی را نشان می‌دهد، مسیرهای متعددی برای بهبود در آینده وجود دارد. یکی از مسیرهای امیدوارکننده، بررسی رویکردهای ترکیبی است که هم تکنیک‌های استخراجی و هم انتزاعی را ترکیب می‌کند. با بهره‌گیری از قوت‌های هر روش، مانند توانایی مدل‌های استخراجی در شناسایی جملات کلیدی و ظرفیت مدل‌های انتزاعی در تولید متن جدید، ممکن است بتوان به خلاصه‌های با کیفیت بالاتر دست یافت. علاوه بر این، بهینه‌سازی و تنظیم دقیق‌تر هایپرپارامترها برای حوزه‌های خاص بیومدیکال می‌تواند منجر به نتایج خلاصه‌سازی هدفمندتر و دقیق‌تر شود. همانطور که حوزه پردازش زبان طبیعی به سرعت در حال پیشرفت است، ادغام تکنیک‌ها و معماری‌های نوظهور ممکن است امکانات جدیدی را برای بهبود خلاصه‌سازی متون بیومدیکال آزاد کند.

مراجع

مراجع

[1] Dalal, Vipul, and Latesh Malik. "A survey of extractive and abstractive text summarization techniques." 2013 6th Internab1. IEEE, 2013.

[2] Bharathi Mohan G; Prasanna Kumar R; Elakkiya R; Siva Jyothi Natha Reddy B; Vemireddy Anvitha " Fine-Tuning Pretrained Transformers for Abstractive News Summarization" 2023 EASCT, IEEE

[3] Hassan Shakil , Ahmad Farooq, Jugal Kalita. "Abstractive text summarization: State of the art, challenges, and improvements"

[4] Tian Shi and Yaser Keneshloo and Naren Ramakrishnan and Chandan, K. Reddy: Neural Abstractive Text Summarization with Sequence-to-Sequence Models, arXiv: (2020)

[5] K. Girthana; S. Swamynathan; A. R. Nirupama; S. Sri Akshya; S. Adhithyan, "Web-based Pretrained Transformer Model for Scientific Paper Summarization (WPT-SPS)"

[6] T.Lin, Y. Wang, X. Liu,and X.Qiu, "A survey of transformers",AI Open,2022.