
بسم الله الرحمن الرحیم

دانشکده مهندسی کامپیوتر

پایان‌نامه کارشناسی ارشد مهندسی هوش مصنوعی

خلاصه‌ساز متون کوتاه فارسی با استفاده از روش‌های مبتنی بر داده

نگارنده: عرفان جلیلی‌جلال

استاد راهنما

دکتر مرضیه رحیمی

اساتید مشاور

دکتر مرتضی زاهدی

ماه و سال

دی ۱۴۰۰

فرم شماره (۳) صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد آقای **عرفان جلیلی جلال** با شماره دانشجویی **۹۸۰۳۷۹۴** رشته **مهندسی کامپیوتر** گرایش **هوش مصنوعی و رباتیکز** تحت عنوان خلاصه‌ساز متون گونه فارسی با استفاده از روش‌های مبتنی بر داده که در تاریخ **۱۴۰۰/۱۰/۱۵** با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار شد به شرح ذیل اعلام می گردد:

الف) درجه عالی: نمره ۲۰-۱۹ ☑ ب) درجه خیلی خوب: نمره ۱۸/۹۹ – ۱۸ ☐

ج) درجه خوب: نمره۱۷/۹۹– ۱۶ ☐ د) درجه متوسط: نمره ۱۵/۹۹-۱۴ ☐

ه) کمتر از ۱۴ غیر قابل قبول و نیاز به دفاع مجدد دارد ☐

نوع تحقیق: نظری ☐ عملی ☐

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱ـ استاد راهنمای اول	مرضیه رحیمی	استادیار	
۳- استاد مشاور	مرتضی زاهدی	استادیار	
۴- استاد داور اول	حمید حسن‌پور	استاد	
۵- استاد داور دوم	هدی مشایخی	استادیار	
۶- نماینده تحصیلات تکمیلی	محسن فرهادی	مربی	

نام و نام خانوادگی رئیس دانشکده:

تاریخ و امضاء و مهر دانشکده:

تقدیم به پدر و مادرم:

خدای را بسی شاکرم که از روی کرم، پدر و مادری فداکار نسیم ساخته تا در سایه درخت پربار وجودشان بیاسایم و از ریشه آنها شاخ و برگ گیرم و از سایه وجودشان در راه کسب علم و دانش تلاش نمایم. والدینی که بودنشان تاج افتخاری است بر سرم و نامشان دلیلی است بر بودنم، چرا که این دو وجود پس از پروردگار، مایه هستی ام بوده اند دستم را گرفتند و راه رفتن را در این وادی زندگی پر از فراز و نشیب آموختند. آموزگارانی که برایم زندگی، بودن و انسان بودن را معنا کردند.

تشکر قلبی و لسانی خود را از استاد عالی سرکار خانم دکتر رحیمی که زحمت راهنمایی این پایان نامه را عهده دار گردیدند و در تمامی مراحل انجام رساله از راهنمایی های مدبّرانه ایشان استفاده نمودم ابراز می دارم و توفیقات روز افزون ایشان را توأم با صحت و سعادت خواستارم...

همچنین از داوران گرامی که زحمت داوری و این پایان نامه را دارند سپاس کمال سپاس را دارم.

خالصانه از تمامی اساتید و معلمان و مدرسانی که در مقاطع مختلف تحصیلی به من علم آموخته و مرا از سرچشمه دانایی سیراب کرده اند

متشکرم.

تعهدنامه

اینجانب **عرفان جلیلی جلال** دانشجوی دوره کارشناسی ارشد رشته هوش مصنوعی و رباتیکز دانشکده کامپیوتر دانشگاه صنعتی شاهرود نویسنده پایان‌نامه **خلاصه‌ساز متون کوتاه فارسی با استفاده از روش‌های مبتنی بر داده** تحت راهنمائی **دکتر رحیمی** متعهد می‌شوم.

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می‌باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان‌نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان‌نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ

امضای دانشجو

چکیده

عنوان‌سازی خودکار یکی از چالش‌های پردازش زبان‌های طبیعی است که در زیرمجموعه خلاصه‌سازی کوتاه قرار می‌گیرد و می‌تواند در کاربردهایی همچون پیشنهاد عنوان برای متون خبری و یا ایمیل کارآمد باشد. در این راستا ماشین سعی می‌کند که خلاصه‌ای بسیار فشرده و کوتاه از اطلاعات متن ورودی را در قالب عنوان به کاربر منتقل کند. برای ساخت عنوان، خلاصه‌سازی چکیده‌ای رویکرد مناسبتری در مقابل رویکرد استخراجی است. در رویکرد چکیده‌ای، برخلاف رویکرد استخراجی که جملات خلاصه را عینا از متن ورودی کپی می‌کند، مدل باید کلمات را به صورت معناداری در کنار هم قرار دهد تا خلاصه‌ای (یا در اینجا به صورت دقیق‌تر عنوانی را) برای متن ورودی تولید کند؛ به همین دلیل است که برای پیاده‌سازی رویکردهای چکیده‌ای نیاز به دادگان عظیم و مدل‌های پیچیده‌ای داریم تا زبان مقصد را درک کرده و عمل خلاصه‌سازی را به صورتی کارا انجام دهد. بنابراین گردآوری دادگان کافی و انتخاب مدل مناسب از چالشهای پیش رو هستند. کارهای ارائه‌شده برای عنوان‌سازهای چکیده‌ای عموما در زبان انگلیسی و دیگر زبان‌های رایج هستند و با استفاده از دادگان عظیم و مدل‌های بازگشتی پیچیده پیاده‌سازی شده‌اند. از آنجایی که زبان فارسی یک زبان با منابع محدود به حساب می‌آید، در پیاده‌سازی یک عنوان‌ساز فارسی به روش چکیده‌ای، چالش‌های مذکور جدی‌تر به نظر می‌رسند. ما در این پایان‌نامه برای فایق آمدن بر مشکل محدودیت دادگان، از شبکه‌ها پیش‌آموزش‌داده‌شده انتقالی استفاده کردیم تا با تعداد دادگان محدودتری عنوان‌ساز فارسی مناسب با دقت بالا به مردم فارسی زبان بصورت آزاد ارائه دهیم. برای اینکار ابتدا تعداد دادگانی مناسب برای اینکار متشکل از حدود ۷۰ هزار جفت چکیده و عنوان متناظر از ژورنال‌های معتبر فارسی زبان با تنوع موضوعی بالا جمع‌آوری کردیم. سپس از مدل زبانی mT5 استفاده کردیم، مدل قبلا با تعداد زیادی متن با حدود ۱۰۰ زبان از جمله فارسی به عنوان یک مدل زبانی آموزش داده شده است که انتظار می‌رود در این مرحله ساختار زبان را دریافت کرده باشد. سپس در ادامه این مدل را برای کاربرد عنوان‌ساز فارسی با دادگان جمع‌آوری‌شده تنظیم کردیم. در این مرحله دقت دادگان تست با معیار ROUGE-1 به مقدار ۴۵ رسید. در ادامه برای بهبود خروجی‌ها و دقت مدل دو معماری پیشنهاد شد. در اولین معماری از یک نویززدا در ادامه‌ی شبکه عنوان‌ساز ارائه شده در مرحله قبلی، اضافه شد تا عنوان‌ها بهبود پیدا کنند. در مدل مذکور، نویززدا همان شبکه mT5 است که این بار عناوین تولیدشده را به عنوان ورودی دریافت کرده و یاد می‌گیرد که عناوین مناسب‌تری را با توجه به عناوین هدف تولید نماید. همچنین در معماری دوم مشابه مدل‌های آنسمبل انتخابی، از چند مدل mT5 استفاده شد تا با استفاده از یک معیار مناسب، از بین عنوان‌های تولید شده بهترین عنوان انتخاب گردد. با این معماری‌های پیشنهادی، دقت را به حدود ۴۶ با معیار ROUGE-1 افزایش دادیم که در مقایسه کارهای مشابه دقت درخور توجهی است.

کلمات کلیدی: یادگیری انتقالی، پردازش زبان‌های طبیعی، خلاصه‌ساز فارسی، عنوان‌ساز فارسی، یادگیری عمیق

لیست مقالات مستخرج از پایان نامه

E. jalili, M. Rahimi **"leveraging transformer models for Persian abstractive title generation"** (submission pending)

فهرست مطالب

فهرست جداول

فهرست اشکال

فصل۱ : مقدمه

خلاصه‌سازی بازنویسی یک نوشته را گویند که مختصرتر باشد درحالی‌که محتوا و ایدهٔ نوشتهٔ مرجع را در برداشته باشد [1] . در گذشته خلاصه‌سازی تنها توسط انسان انجام می‌شده اما اکنون با وجود الگوریتم‌های یادگیری ماشین و مفاهیم پردازش زبان‌های طبیعی این کار توسط کامپیوتر هم قابل انجام شده است. یکی از کارهای بسیار مشابه به خلاصه‌سازها در پردازش زبان‌های طبیعی، مولد عنوان است که به دلیل وجود شباهت بالا با خلاصه‌ساز خودکار معمولا این دو را در یک دسته قرار می‌دهند [2]. مولد عنوان چکیده‌ای باید متن ورودی را پردازش کند و در خروجی کلمات را طوری کنار هم بچیند که بتواند همچون عنوانی مناسب برای متن مرجع در نظر گرفته شود. یک عنوان خوب باید علاوه بر کوتاه بودن، موضوع اصلی متن و اطلاعاتی که قرار است خواننده با آن مواجه شود را بازتاب دهد. از کاربردها مولد عنوان می‌توان به پیشنهاد خودکار عنوان، برای متن ایمیل، اخبار و مقالات اشاره کرد. مشابه خلاصه‌سازهای چکیده‌ای، برای پیاده‌سازی این ابزار مفید نیاز به دادگان مناسب و کافی وجود دارد. از آنجایی که تا بحال برای زبان فارسی مولد چکیده‌ای آزاد ارائه نشده است سعی شد در این پایان‌نامه تحقیقاتی جهت ارائه‌ی یک عنوان‌ساز فارسی مناسب برای فارسی زبانان ارائه شود.

از ظهور یادگیری ماشین تاکنون، خلاصه‌سازی خودکار یکی از چالش‌های پردازش زبان‌های طبیعی بوده است. برای خلاصه‌سازی دو رویکرد استخراجی و چکیده‌ای وجود دارد. در رویکرد استخراجی[1] جملات خروجی از متن مرجع به صورت مستقیم انتخاب می‌شوند. در رویکرد چکیده‌ای[2]، خلاصه‌ساز با قرار دادن کلمات مناسب جمله‌سازی می‌کند و خلاصه نهایی را ارائه می‌دهد. رویکرد چیکده‌ای به خلاصه‌ای که توسط انسان نوشته می‌شود نزدیک‌تر است به همین دلیل بیشتر مورد توجه محققان قرار گرفته است [3]. خلاصه‌سازی چکیده‌ای کار پیچیده‌ای به حساب می‌آید چرا که برای اینکار نیاز به درک از ساختار زبان و نحوه‌ی خلاصه‌سازی است. با پیشرفت علم یادگیری ماشین و ظهور یادگیری عمیق روزبه‌روز خلاصه‌سازهای چیکده‌ای دقیق‌تر می‌شود. با یادگیری ماشین می‌توان مدلی را به صورت باناظر یا خودناظر آموزش داد که

[1] Extractive
[2] Abstractive

علاوه بر درک دستور زبان مقصد و روابط معنایی، نحوه خلاصه‌سازی هم یاد بگیرد [4]. همانطور که گفته شد برای این مدل‌ها نیاز به مجموعه دادگانی وجود دارد تا آموزش به روش باناظر یا خودناظر انجام شود، به همین جهت این روش‌ها را مبتنی برداده[1] هم می‌نامند. لازم به ذکر است برای رویکرد چیکده‌ای روش‌های مبتنی بر غیر یادگیری ماشین مانند درخت‌های معنایی هم ارائه شده است [4-6]. اما عمده تحقیقات بروز با توجه به وجود داده های متنی، با روش‌های مبتنی بر داده انجام می شود [4].

همانطور که گفته شد برای خلاصه‌سازی چکیده‌ای نیاز به درک عمیقی از زبان مقصد است و اینکار تنها با یادگیری‌ماشین قابل تصور است. شبکه‌ها بازگشتی[2] یکی از ابزارهای یادگیری ماشین است که برای داده‌ها دنباله‌ی[3] ارائه شده‌اند. این شبکه‌ها دارای حافظه هستند پس می‌توانند الگوها و روابط اعضای دنباله را شناسایی کنند. با یک متن می‌توان مانند یک دنباله رفتار کرد؛ اعضای یک دنباله‌ی متنی کلمه‌ها می‌باشند که روابط زبانی در بین آن‌ها حاکم است. با این وجود، شبکه‌های بازگشتی می‌تواند پاسخ خوبی برای پیاده‌سازی خلاصه‌سازها باشند [4]. همانطور که اشاره شد، شبکه‌ی بازگشتی علاوه بر یادگیری قواعد دستوری و معنایی متن باید یاد بگیرد چگونه خلاصه‌ای تولید کند تا اطلاعات اصلی متن مرجع را در بر داشته باشد؛ به همین دلیل است که برای آموزش اینگونه مدل‌ها از ابتدا نیاز به دادگان عظیم مناسب برای خلاصه‌سازی وجود دارد که برای زبان‌هایی با منابع محدود مانند فارسی ناممکن و هزینه‌بر است [2]. در سال‌های اخیر خلاصه‌سازهای چکیده با استفاده از مدل‌هایی با معماری دنباله‌به‌دنباله بر پایه‌ی شبکه‌های بازگشتی LSTM ارائه شدند که عمدتا بر روی زبان انگلیسی با مجموعه دادگاه عظیم مانند CNN-DailyNews [7] با ۱.۳ میلیون جفت نمونه‌های متن و خلاصه‌ی آن، آموزش داده شده اند [8]. در صورتی که برای زبان فارسی، که یک زبان منابع محدود به حساب می‌آید [9]، جمع‌آوری داده‌های کافی برای آموزش مدل‌های دنباله‌به‌دنباله از ابتدا، هزینه‌بر و ناکارآمد است. علاوه بر اینکه برای بکارگیری از شبکه‌های

[1] Data-driven
[2] Recurrent Neural Networks
[3] Sequential

بازگشتی نیاز به دادگان عظیم وجود دارد، این شبکه‌ها دارای اشکلاتی همچون تولید کلمات و عبارات تکراری در خروجی و مهو شدن گرادیان با افزایش طول داده‌ها هستند. همچنین در این شبکه‌ها روابط معنایی و ارتباطات بین کلمات به خوبی استخراج نمی‌شوند [10]. این چالش‌ها برای زبان فارسی بیشتر مشکل‌ساز هستند چرا برای اعمالی مانند خلاصه‌سازی با شبکه‌ها بازگشتی دادگان کافی وجود ندارد به همین خاطر است که تا به الان خلاصه‌ساز مناسبی برای زبان فارسی با استفاد از شبکه‌ها بازگشتی ارائه نشده است.

در ادامه‌ی تحقیقات بر روی شبکه‌های بازگشتی، با ظهور مدل‌های انتقالی[1] [11] تحولی دیگر در پردازش زبان‌های طبیعی ایجاد شد [12]. مدل‌های زبانی انتقالی پیش‌آموزش دیده‌شده با استفاده از مجموعه متون عظیم با روش‌های خودناظر آموزش داده شده‌اند. به عنوان مثال مدل‌زبانی از پیش‌آموزش GPT2 [13] متشکل از ۱.۵ بیلیون پارامتر می‌باشد، با ۸ میلیون صفحه وب به روش خودناظر آموزش دیده است. با ارائه این مدل‌ها توسط کمپانی‌های بزرگ، نیاز به دادگان عظیم و منابع محاسباتی برای کارهای پایین‌دستی[2] مانند خلاصه‌سازی اتوماتیک یا ترجمه اتوماتیک کم‌رنگ‌تر می‌شوند. چرا که این مدل‌ها از قبل ساختار زبان را درک کرده‌اند و برای پیاده‌سازی کارهای پایین‌دستی مانند خلاصه‌سازی تنها نیاز است آن‌ها را برای کار مورد نظر تنظیم کرد [14]. با رشد این مدل‌های زبانی، راه برای اعمال پردازش زبان طبیعی برای زبان‌هایی با منابع محدود و همچنین محققانی با منابع پردازشی به مراتب کمتر، میسرتر شد. این پیشرفت برای زبان زبان فارسی خبر خوبی بود چرا که می‌توان با دادگان بسیار کمتری اعمالی که درگذشته نیاز به دادگان کمتری داشته‌اند را آموزش داد و مورد بهره‌گیری قرار داد.

در جهت پیاده‌سازی عنوان‌ساز فارسی ما تصمیم گرفتیم برای رفع چالش‌های رویکردها گذشته با بهره‌گیری از مدل‌های زبانی انتقالی پیش‌آموزش‌داده‌شده یک عنوان‌ساز برای زبان فارسی با دقت مناسب ارائه کنیم. برای اینکار از مدل زبانی پیش‌آموزش داده شده‌ی mT5 بهره گرفتیم چرا که این مدل چند

[1] Transformer
[2] Down-stream tasks

زبانه است و علاوه بر انگلیسی و زبان‌های رایج دیگر، با زبان فارسی هم به عنوان یک مدل زبانی توسط شرکت گوگل آموزش دیده شده است. در این مدل از مکانیزم توجه [11] بهره گرفته‌شده که به خوبی می‌تواند روابط زبانی بین کلمات را برای کار مورد نظر یاد بگیرد. برای آموزش این مدل متادیتای حدود هفتاد هزار مقاله از ژونال‌های معتبر فارسی، با موضوعات مختلف بارگذاری شد. برای اینکار ما از چکیده و عنوان مقالات بارگذاری شده برای آموزش عنوان‌ساز فارسی استفاده کردیم. پس از آموزش مدل mT5 با دادگان جمع‌آوری شده، دقت عنوان‌ساز آموزش‌داده‌شده با معیار ROUGE به عدد حدود ۴۵ رسید که نسبت به خلاصه‌ساز ارائه شده فارسی و کارهای مشابه انگلیسی این دقت قابل توجه است.

فصل ۲ : مرور کارهای پیشین

۲-۱ مقدمه

با رشد روزافزون اطلاعات متنی دیجیتال در دهه هفتاد میلادی، ارائه نتایج مرتبط برای جست‌وجوی کاربر اهمیت فزاینده‌ای پیدا کرده بود. این تحقیقات در حوزهٔ بازیابی اطلاعات[1] (به‌اختصار IR) انجام می‌شد. به‌عنوان‌مثال تحقیقاتی جهت رتبه‌بندی متون در راستای مرتبط بودن با جست‌وجوی کاربر، انجام شده است. در این تحقیقات سعی بر این بود که متون بر اساس ارتباط آن‌ها با اطلاعاتی که کاربر جست‌وجو می‌کند، رتبه‌بندی شوند [15,16]. از دیگر تحقیقات مهم در بازیابی‌اطلاعات ، می‌توان به مرجع [17] اشاره کرد که با استفاده از شناسایی قواعد گرامری[2] سازوکاری برای خودکارسازی تولید لیست کلمات مهم[3] برای یک کتاب به همراه آدرس رجوع به آن کلمه در متن، پیشنهاد داده است.

در ادامهٔ تحقیقات حوزهٔ بازیابی اطلاعات، با ظهور الگوریتم‌های یادگیری ماشین و پردازش زبان‌های طبیعی، همواره تحقیقاتی در جهت توسعه خلاصه‌ساز اتوماتیک هوشمندی انجام شده تا بتواند ایدهٔ متن اصلی را در قالب مختصرتری ارائه دهد. دو رویکرد بنیادی استخراجی و چکیده‌ای برای خلاصه‌سازی وجود دارد؛ در ادامه به کارهای انجام شده در هر یک خواهیم پرداخت.

۲-۲ خلاصه‌ساز استخراجی

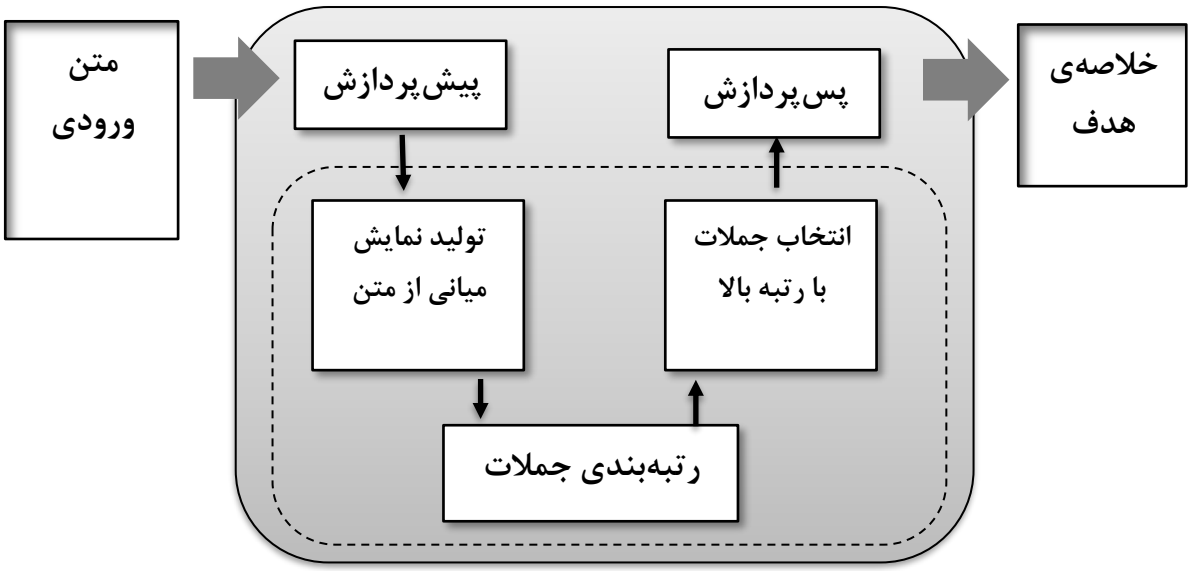
تحقیقات انجام شده بر روی خلاصه‌سازی خودکار در دههٔ نود میلادی و دههٔ اول قرن حاضر اغلب با رویکرد استخراجی بوده است. چند رویکرد آماری، معنایی و غیره برای این روش از خلاصه‌سازی ارائه شده است. نمای کلی خلاصه‌سازی استخراجی در نمای کلی از معماری خلاصه‌سازهای استخراجی آورده شده است. در مرحله اول و آخر، به ترتیب پیش‌پردازش داده‌ها و پس‌پردازش آن‌ها انجام می‌شود. در شکل ۲–۱ مرحله

[1] Information retrival
[2] Syntatic analysis
[3] Book indexing

پیش‌پردازش، داده‌های خام مورد پردازش‌هایی مانند قطعه‌بندی[1] و حذف علائم خاص قرار می‌گیرند و تغییراتی بر روی آن‌ها اعمال می‌شود. و پس از انجام پردازش در نهایت پس‌پردازش انجام می‌شود؛ در این مرحله اعمالی مانند جابه‌جایی جملات و در صورت نیاز کم کردن جملات برگزیده انجام می‌شود. در مرحله پردازش اصلی متون‌های مختلفی برای اینکار ارائه شده است. به صورتی عمومی می‌توان مراحل پردازش اصلی به ترتیب زیر است:

- تبدیل متن به یک نمایش[2] مناسب مانند bag-of-words تا بتوان بر روی متن آنالیز انجام داد [18].
- رتبه‌بندی جملات بر اساس نمایشی که برای دادها در مرحله قبل در نظر گرفته شده است [19].
- استخراج جملاتی که از همه مهم‌تر هستند بر اساس رتبه‌بندی که در مرحلهٔ قبل انجام می‌شود [3]

[1] Tokenization
[2] Represantation



شکل ۲–۱ نمای کلی از معماری خلاصه‌سازهای استخراجی مرجع [3]. خلاصه‌سازهای استخراجی دارای مراحل پیش‌پردازش، پردازش و پس‌پردازش هستند.

تفاوت عمدهٔ خلاصه‌ساز استخراجی در روش رتبه‌بندی جملات مهم و گزینش آن‌های برای متن خروجی است. روش‌ها مختلفی برای روش استخراجی از جمله آماری، معنایی، قطعه‌ای[1]، گرافی و غیره ارائه شده‌اند [3]. در ادامه به شرح بعضی از این روش‌ها خواهیم پرداخت.

در متدهای آماری بیشتر به تحلیل آماری کلمات پرداخته می‌شود. تعداد وقوع کلمات و محل وقوع کلمه، یکی از فاکتورهای رتبه‌بندی جملات برای تولید خلاصه نهایی می‌باشد. در مقاله [20] فرض شده است که آن کلماتی که بیشتر تعداد وقوع در متن را دارند بار معنایی بیشتری را منتقل می‌کنند تا کلماتی که کمتر تکرار شده‌اند؛ به توجه به این فرض با استفاده از روش TF-IDF [21] هر جمله استخراج ویژگی شده است. در نهایت با استفاده روش بدون‌ناظر[2] k-mean مشابه مرجع [22] جملات مشابه به دسته‌های جدا تقسیم می‌شوند در نهایت از هر دسته آن جمله‌ای که به مرکز نزدیک‌تر است برای خلاصه خروجی برگزیده می‌شود. در مرجع [23]، که در فارسی انجام شده است؛ برای هر جمله یازده ویژگی استخراج می‌شود. در ادامه،

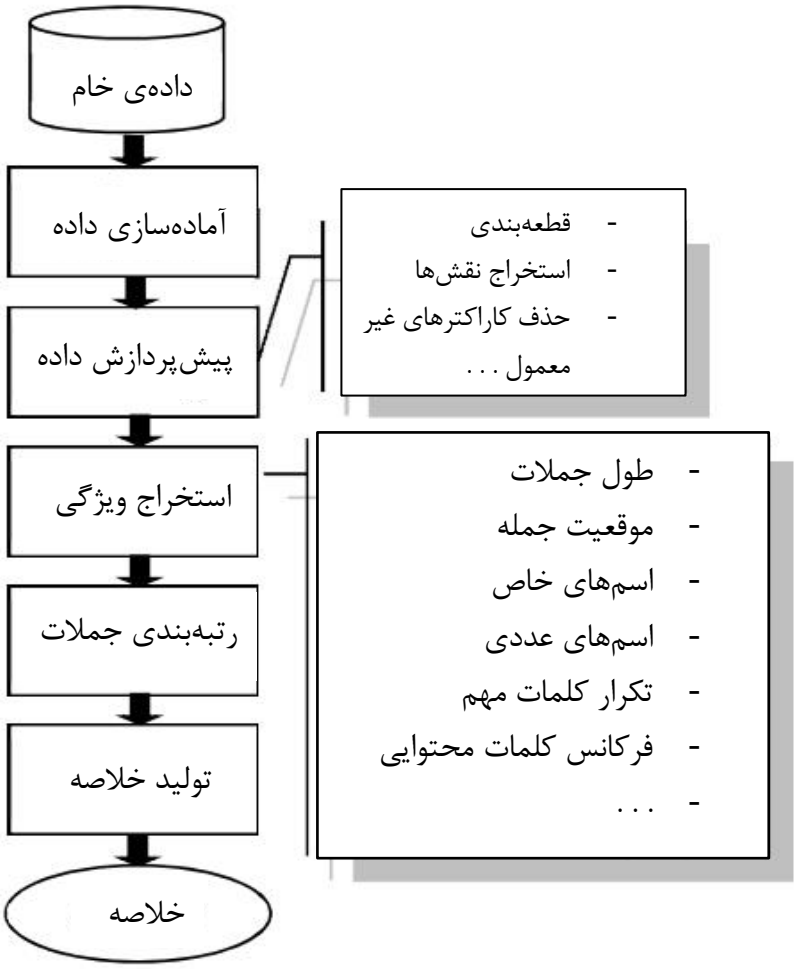
[1] Token
[2] Unsupervised

مواردی از آن‌ها را توضیح خواهیم داد:

- اسم خاص: جملاتی که دارای تعداد بیشتری اسم خاص باشند، اهمیت بیشتری دارند.
- طول جملات: آن جملات بلندتر مهم‌تر فرض شده است.
- داده‌ها عددی: وجود داده‌های عددی و تعداد آن‌ها بر اهمیت جمله می‌افزاید.
- موقعیت جمله: آن جملاتی که در ابتدا و انتهای متن ورودی قرار گرفته‌اند بااهمیت‌تر هستند.
- فرکانس کلمات محتوایی: برای این ویژگی فرکانس کلمات محتوایی (فعل، فاعل، مفعول و غیره که مفهوم جمله را در بردارند) محاسبه شده است.

در نهایت مقادیر هر یک از ویژگی‌ها در ضریب خاص خود (که اهمیت آن ویژگی را نشان می‌دهد) ضرب می‌شود و با یکدیگر جمع می‌شوند؛ به این شکل رتبهٔ نهایی هر جمله محاسبه می‌شود. در آخر، باتوجه‌به طول خلاصهٔ موردنظر جملات با رتبه‌بندی بیشتر انتخاب می‌شوند. با استفاده از این متد با استفاده معیار ROUGE-1 عدد ۳۶ کسب کرده است. در شکل ۲–۲ نمایی از معماری پیشنهاد شده آورده شده است.

از مزایای روش‌های آماری می‌توان به نیاز محدود به منابع محاسباتی و حافظهٔ اشاره کرد [۲۴]. همچنین برای پیاده‌سازی این نوع خلاصه‌ساز نیاز به آگاهی از منابع زبان‌شناسی نمی‌باشد [۲]. همچنین از معایب این رویکرد می‌توان به رتبه‌بندی ناعادلانه جملاتی با مفاهیم تکراری یا غیر مهم اشاره کرد؛ به این شکل که ممکن است، این جملات رنک بالاتری نسبت به جملات مهم‌تر بگیرند [۲۵].

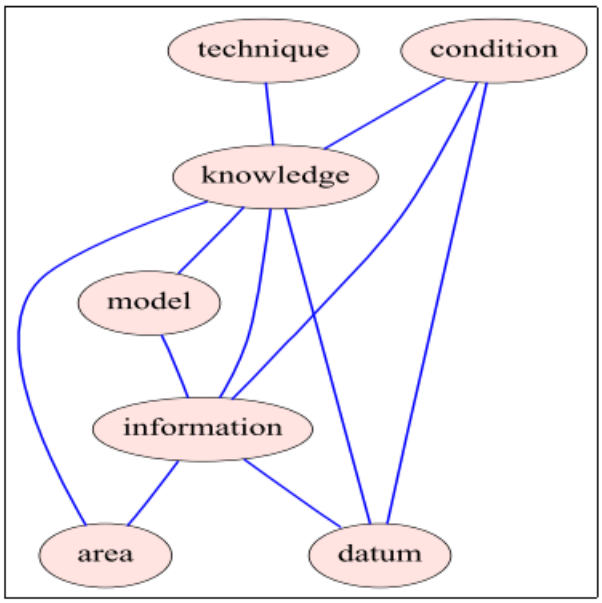


شکل ۲–۲ معماری پیشنهاد شده برای خلاصه‌سازی استخراجی، منبع [23].

همان‌طور که گفته شد در خلاصه‌سازی استخراجی چالش اصلی پیداکردن یک نمایش[1] برای جملات و رتبه‌بندی هر جمله و در نهایت گزینش جمله‌ها بر اساس رتبه‌بندی است. روش‌های معنایی[2] از دیگر روش‌هایی است که در ادامه روش آماری مورد توجه قرار گرفتند؛ در این دسته از رویکردها، برای نمایش جملات از روش‌های مفهومی و معنایی برای نمایش داده‌ها استفاده می‌شود. به‌عنوان‌مثال، در مرجع [24] با استفاده از دیتابیس‌های واژگانی WordNet [25]، زنجیره واژگانی[3] جملات به عنوان نمایشی از آن

[1] Representation
[2] Semantical methods
[3] Lexical chain

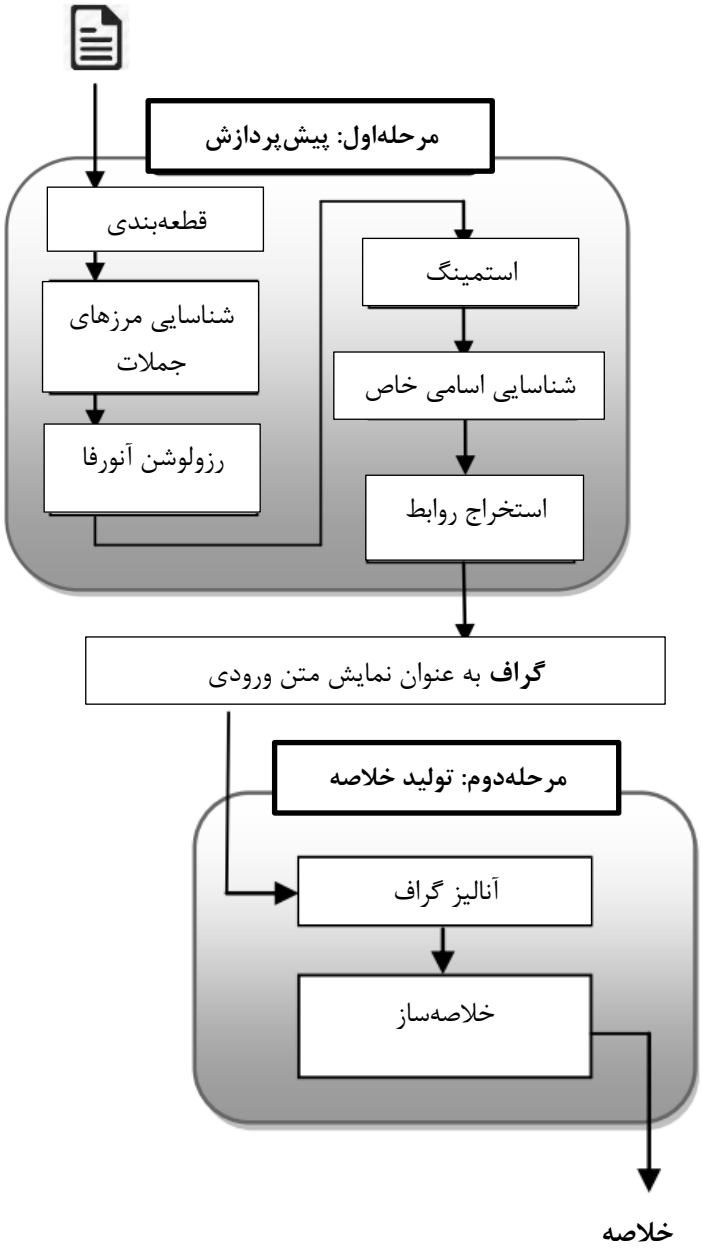
جملات ساخته می‌شود. با این کار انتظار داریم معنای جملات برای نمایش آن‌ها در مراحل بعدی در نظر گرفته شده باشند. در اینصورت جملاتی که مفاهیم و معانی عمیق‌تری را در برداشته، می‌تواند انتخاب مناسبی برای خلاصه نهایی باشند. بعد از استفاده از الگوریتم زنجیره واژگانی برای تولید نمایشی از جملات، نوبت به رنک‌بندی جملات با استفاده از نمایش‌ها می‌باشد. در این کار با استفاده از معیارهای مانند طول زنجیره، جملات رنک‌بندی می‌شوند. درشکل ۲–۳ نمایی از نمایش متون برای خلاصه‌سازی استخراجی در مرجع [24] آورده شده است.



شکل ۲–۳ نمایی از نمایش متون برای خلاصه‌سازی استخراجی در مرجع [24] با استفاده از نظریه گراف‌ها. در این شکل نمایشی از ارتباطات واژگانی بین چند کلمه با استفاده از نمایش گراف‌ها ارائه شده است.

در مرجع [26] برای زبان فارسی سعی شده‌است با استفاده از یک روش مبتنی بر معنا مهم‌ترین جملات بر اساس اهمیت آن‌ها از نظر معنای برای خلاصه خروجی انتخاب شود. برای این کار از پیکره‌گان FarsNet [27] و نظریه گراف‌ها استفاده شده تا نمایش از متن ورودی متنی بر معنی برای خلاصه‌سازی ارائه شود. در این کار، ابتدا در کنار پیش‌پردازش‌های دیگر عملیات NER [28] انجام می‌شود. در این مرحله از پیش‌پردازش اسم‌های مهم مانند اسم‌های خاص، نام افراد، تاریخ و غیره شناسایی می‌شوند. در ادامه این

اسم‌های مهم، به عنوان رئوس گراف در نظر گرفته می‌شود و یال‌های از روابط معنایی که از FarsNet استخراج می‌شود رئوس را به هم وصل می‌کنند. در واقع یال‌ها ارتباط معنایی دو راس را نشان می‌دهند. در ادامه با استفاده از معیارهای پیشنهادی کلمات مرکزی و در ادامه جملات مرکزی به عنوان کاندیدهایی برای خلاصه نهایی انتخاب می‌شوند. در شکل ۲–۴ نمایی از ساختار مدل ارائه شده در مرجع [26] آورده شده است.



شکل ۲–۴ نمایی از معماری ارائه شده در مرجع [26]. شکل بالا مختص به مرحله پیش‌پردازش و شکل پایین مربوط به مرحله پس‌پردازش و تولید خلاصه‌ساز است.

در مرجع [24] و [26] از روشی مبتنی بر معنی به کمک نظریه‌ی گراف‌ها پیشنهاد شده است. از مزایای این روش می توان به موارد زیر اشاره کرد:

- روشی مستقل از زبان است و می‌تواند برای زبان‌های دیگر به کمک دیتابیس‌های واژگانی مانند WordNet پیاده‌سازی شود [29].

- خروجی خلاصه روان‌تر و خواناتر است و می‌تواند اطلاعات اضافی را حذف کند [30].

همچنین معایب این روش از قرار زیرند:

- ممکن است در خروجی مساله دنگلینگ آنفورا[1] پیش بیاید [29].
- معیار شباهت برای ارزیابی شباهت معنایی جملات در مرحله انتخاب، به‌خوبی نمی‌تواند معنی را در نظر بگیرد و ممکن است این مرحله به‌خوبی انجام نشود [30].

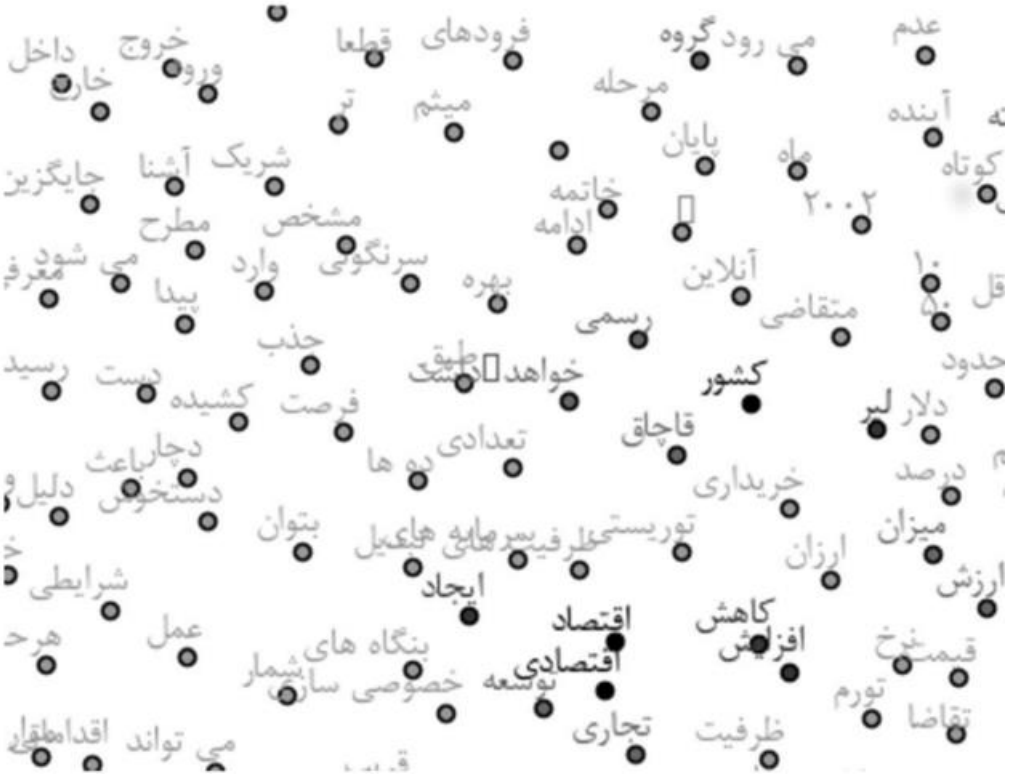
از دیگر رویکردهای خلاصه‌سازی استخراجی می‌توان به روش‌های مبتنی بر یادگیری ناظر اشاره کرد در ادامه به دو مرجع با این روش برای زبان انگلیسی و فارسی اشاره خواهیم کرد.

در مرجع [22] با استفاده از یادگیری با نظارت [2] مدلی آموزش داده شده است که جملات مهم متن ورودی را استخراج کنید. بدیهی است برای آموزش مدل نیاز به دادگانی شامل جفت متن و خلاصه متن می‌باشد. در این مقاله ابتدا استخراج ویژگی با روش‌هایی آماری و مبتنی بر فرکانس کلمات مانند TF-IDF [31] انجام می‌شود. همچنین در ادامه برای انتخاب جملات مهم با خلاصه نهایی یک مدل بیزین برای رنک‌بندی جملات مهم آموزش داده می‌شود. یکی از معایبی که این مرجع دارد استفاده از روش‌های آماری برای استخراج ویژگی است که در مواردی نمی‌تواند به خوبی معنا را در بربگیرد. از مزایای این روش می‌توان به سرعت بالای پردازش اشاره کرد [3].

از کارهای اخیر با رویکرد استخراجی در زبان فارسی می‌توان به مقاله [32] اشاره کرد. در این کار ابتدا با استفاده از پیکره‌گان[3] همشهری۲ [33] جاسازی‌های[4] مناسب برای کلمات تعیین می‌شود؛ با اینکار هرکلمه فارسی (مجموعه کلماتی که در این مجموعه نوشته وجود دارد) را به یک بردار عددی نگاشت داده می‌شود. در این شیوه از نمایش، کلماتی که معانی نزدیکی به هم دارند، فاصله برداری کمتری نیز دارند. در

[1] Dangling Anaphora
[2] Supervised Learning
[3] Corpus
[4] Embedding

این کار بعد از تولید جاسازی‌های کلمه‌های رایج فارسی، برای شناسایی موجودیت‌نام‌گذاری‌شده[1] (NER)، یک شبکه عصبی آموزش داده می‌شود. در نهایت جملات بر اساس تعداد کلمه‌های NER برای خلاصه نهایی رنک‌بندی می‌شوند. در شکل ۲–۵ نمایی از جاسازی بدست آمده با استفاده از پیکره‌گان همشهری۲ آورده شده است. برای اینکار ابتدا باید ابعاد بردار جاسازی به سه بعد یا کمتر کاهش داد تا بتوان آن‌ها را رسم کرد. همانطور که مشاهده می‌شود کلماتی که معانی نزدیک‌تری دارند در کنار هم قرار گرفته‌اند؛ مانند کلمه‌ها «خرید» و «ارزان» و یا کلمات «افزایش»، «کاهش» و «ظرفیت»؛ این رفتار نشان دهنده‌ی توجه به معنی کلمه در نمایش آن است [32].



شکل ۲–۵ نمونه‌ای از نمایش جاسازی مرجع [32]، با استفاده از پیکرده‌گان همشهری۲. مشاهده می‌شود کلماتی با معانی نزدیک در کنار هم قرار گرفته‌اند.

از مزایای روش پیشنهادی در مرجع [32]، استفاده از جاسازی به عنوان نمایشی از متن است. با این رویکرد

[1] Named Entity Recignition

بهتر می‌توان معنا را درنظر گرفت. از معایب این روش می‌توان وجود نیاز به داده‌های برچسب‌گذاری شده اشاره کرد [3].

از مزایای رویکرد استخراجی برای خلاصه‌سازی می‌توان به موارد زیر اشاره کرد.

- این رویکرد ساده‌تر و سریع‌تر است و برای پیاده‌سازی آن نیاز به منابع محاسباتی بالا نمی‌باشد [3].
- باتوجه‌به اینکه جملات مستقیم از ورودی مرجع انتخاب می‌شود، در نتیجه دقت مناسبی را با معیارهای رایج خلاصه‌سازی دریافت می‌کنند. چرا که این معیارهای عموماً بر پایهٔ وجود یا عدم وجود کلمهٔ یا دنباله کلمه‌های متن مرجع در خلاصهٔ نهایی می‌باشند [34].

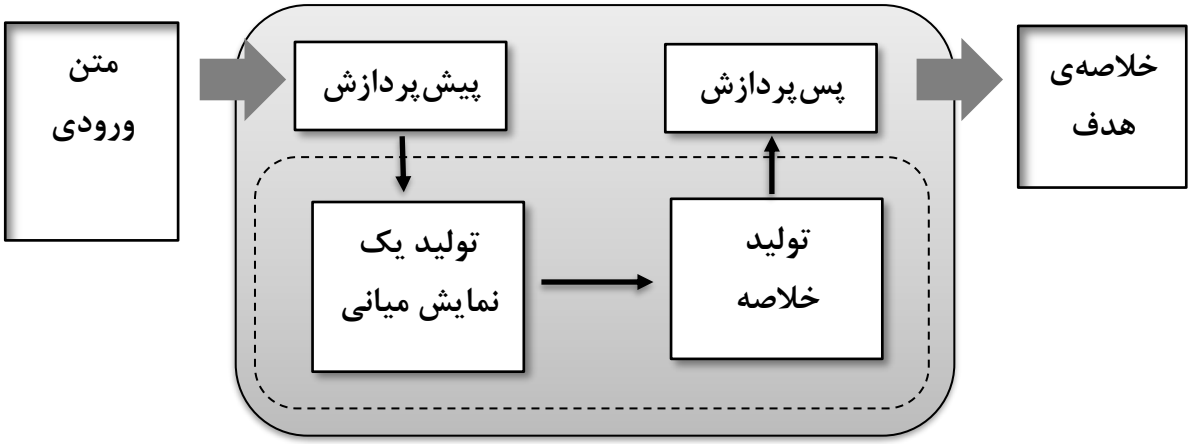
از معایب روش استخراجی می‌توان به

- عدم روان بودن و انسجام در خلاصه‌های خروجی احساس می‌شود. به دلیل اتصال نابجای بعضی جملات، روانی جملات کاهش پیدا می‌کند [6].
- شباهت کمی به خلاصه‌های دست‌نویس توسط انسان وجود دارد [3].
- ممکن است در جملات خروجی افزونگی معنایی وجود داشته باشد [35].
- احتمال بروز مشکل دنگلینگ آنفورا در این رویکرد وجود دارد [6].

۲-۳ خلاصه‌ساز چکیده‌ای

در رویکرد چکیده‌ای بر خلاف رویکرد استخراجی، خلاصه‌ساز متن ورودی را مستقیما رونوشت نمی‌کند بلکه کلمات را به صورت معنادار با قواعد زبان مرجع، طوری در کنار هم می‌چیند که مفهوم اصلی متن مرجع را به صورت خلاصه در خروجی ارائه دهد [۳۶]. به صورت کلی برای رویکرد چیکده‌ای نیاز به درک بیشتری از قواعد زبانی و معنایی متن می‌باشد [۳۷]. با استفاده از دادگان مناسب به همراه روش‌های باناظر می‌توان مدلی آموزش داد که علاوه بر دریافت ساختارهای زبانی، بتواند خلاصه‌ای از متن اصلی تولید کند.

البته برای این رویکرد روش‌های مبتنی بر غیر روش‌های یادگیری ماشین هر ارائه شده است [۹]. به صورت کلی ساختارهای خلاصه‌سازهای چکیده‌ای شکل ۲–۶ نمایش داده شده است.



شکل ۲–۶ معماری کلی خلاصه‌سازهای چکیده‌ای، مرجع [3].

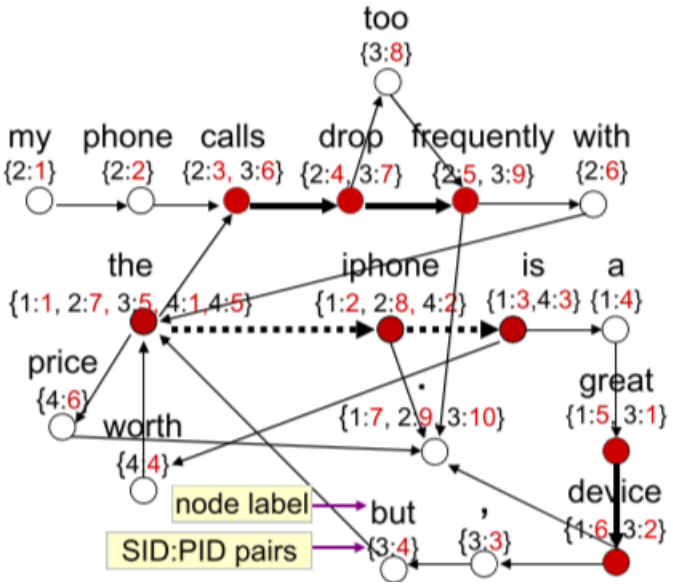
مراحل نشان داده شده در شکل ۲–۶ به شرح زیر است [۲]:

- **پیش‌پردازش** در این مرحله داده‌ها بسته به نوع مدل، قبل از ورودی‌دادن، مورد پیش‌پردازش قرار می‌گیرند.
- **پردازش** این بخش مهم‌ترین بخش خلاصه‌ساز است و به ترتیب زیر انجام می‌شود:
  - متن به یک نمایش برای پردازش تبدیل می‌شود. به‌عنوان‌مثال از رمزنگار برای گرفتن بردار عددی به‌عنوان نمایشی از متن استفاده می‌شود. طوری که روابط معنایی استخراج شوند [4].
  - با گرفتن نمایشی از متن از مرحله قبل، پردازش بر روی آن انجام می‌شود تا در ادامه خلاصه تولید شود. پس‌پردازش در بخش خروجی رمزگشا بسته به نوع مدل مورد پردازش قرار می‌گیرد در نهایت متن خلاصه در خروجی تولید می‌شود. یکی از پردازش‌های که معمولاً انجام می‌شود تبدیل بردار عددی خروجی، به کلمات زبان

مقصد است.

در ادامه به بررسی کارهای مهم انجام شده با این رویکرد در دو زبان فارسی و انگلیسی می‌پردازیم.

در مرجع [۵] یک روش بر پایه نظریه گراف‌ها برای خلاصه‌سازی چکیده‌ای با نام «Opiosis» پیشنهاد شده است. برای نمایش متن مرجع، گره‌های گراف نماینده‌ای از هر کلمه هستند و همچنین بر اساس موقعیت کلمه‌ها در متن مرجع، گره‌ها به یکدیگر متصل می‌شوند. در واقع یال‌ها ساختار جمله‌ها را نمایش می‌دهند. در شکل ۲–۶ نمایی از نحوه نمایش متون در این مرجع آورده شده‌است.



شکل ۲–۷ نمایش پیشنهاد شده برای متن ورودی بر پایه‌ی گراف مرجع [36].

در مقاله [36] مراحل انجام خلاصه‌سازی به ترتیب:

- تشکیل گراف متنی (با روش پیشنهاد داده شده) به عنوانی نمایشی از متن اصلی.
- تولید خلاصه: در این مرحله زیرمسیرها[1] از گراف انتخاب می‌شوند و بصورت زیر رتبه‌بندی شده و برای خلاصه انتخاب می‌شوند:
  - تمام مسیرهای رتبه‌بندی شده به‌صورت نزولی چیده می‌شوند.

[1] Subpath

- مسیرهایی که شباهت زیادی باهم دارند با استفاده از معیارهای شباهتی مانند جکارد[1] حذف می‌شوند.
- با توجه طول خلاصهٔ نهایی که توسط کاربر تعیین می‌شود؛ چند جمله با رنک بالا انتخاب می‌شود.

از مزایای این کار، افزونگی[2] کمتر در خلاصه‌ی خروجی است; در نتیجه این روش می‌تواند مناسب داده‌هایی باشند که تکرار و افزونگی معنا در آن زیاد باشد. به عنوان‌مثال، نظرات کاربران درباره محصولات و داده‌های متنی مشابه می‌تواند برای این مدل برای تولید خلاصه‌ای بدون افزونگی کارا باشد [36].

در مرجع [36] سعی شد تا با تحلیلی نحوی و استفاده گراف نمایشی از متن ارائه شود. مشابه اینکار در مرجع [6]، از نظر استخراج ساختارهای نحوی و عبارت‌های زبانی، از درخت تجزیه وابستگی[3] برای نمایش متن استفاده کرده است. در این مرجع و کارهای مشابه به صورت کلی روند زیر برای تولید خلاصه چکیده‌ای انجام می‌شود:

۱. متن مورد تجزیه قرار داده می‌شود تا تمام درخت‌های وابستگی نحوی استخراج شوند.

۲. درخت‌های وابستگی جزئی[4] از درخت‌های وابستگی نحوی استخراج می‌شوند.

۳. خوشه‌بندی بر اعضای استخراج شده در مرحله قبلی انجام می‌شود تا اطمینان حاصل شود که گوناگونی محتوا در نظر گرفته شده است.

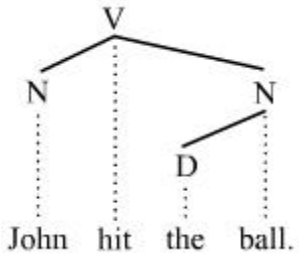
۴. از هر خوشه یک درخت جزئی انتخاب می‌شود که یک جمله را نمایش می‌دهد و نماینده‌ای از کل آن خوشه برای خلاصه نهایی است.

---

[1] Jaccard
[2] redundancy
[3] dependency-based parse tree
[4] Partial dependency trees

۵. با تکنیک‌های مانند خطی‌سازی[1] درخت جزئی به جمله تبدیل می‌شود.

شکل ۲–۸ درخت تجزیه وابستگی برای یک جمله رسم شده است. همانطور که گفته شد از این نمایش برای مرجع [6] استفاده شده است.



شکل ۲–۸ درخت تجزیه وابستگی برای جمله‌ی "John hit the ball"

ازآنجایی‌که خلاصه‌سازی چکیده‌ای مستلزم تحلیل معنایی عمیق‌تری از متن است [3]. در مرجع [5] خلاصه‌ساز چیکده‌ای چندمتنه ارائه شده است. به اینصورت که، از برچسب‌گذاری قواعد معنایی[2] یا SLR استفاده می‌شود تا ساختار نهاد گزاره[3] از جملات موجود در مجموعه متون استخراج شود. در این مرجع مراحل زیر را برای خلاصه‌سازی چندمتنه چکیده‌ای به قرار زیر است:

۱. متون ورودی با ساختار نهاد گزاره و با استفاده از SLR نمایش داده می‌شود.

۲. جملات مشابه در این نمایش خوشه‌بندی می‌شوند.

۳. جملات با استفاده از ویژگی‌های وزن‌دار رتبه‌بندی می‌شود و وزن‌ها با استفاده از الگوریتم ژنتیک بهینه می‌شوند.

۴. با استفاده از مدل‌های تولید زبان با نمایش‌ها نهاد گزاره جملات تولید می‌شوند.

از مزایای استفاده از خلاصه‌سازها بر پایه قواعد نحوی می‌توان به خروجی روان و منسجم اشاره کرد. همچنین از معایب این رویکرد توجه به قواعد نحوی بیشتر از قواعد معنایی است و وابستگی آن به

[1] Linearization
[2] Semantic Role Labeling (SLR)
[3] Predicate argument structures

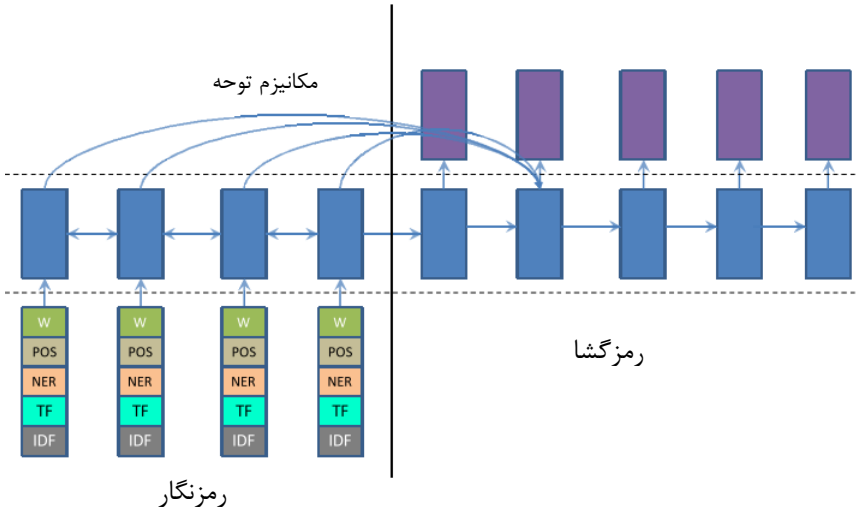
تجزیه‌کننده‌های[1]های موجود را می‌توان اشاره کرد.

از دیگر رویکردها برای پیاده‌سازی خلاصه‌ساز چکیده‌ای استفاده از مفاهیم یادگیری عمیق است [3]. یادگیری عمیق با مفهوم یادگیری مدلی که علاوه بر پیش‌بینی خروجی مورد نظر، بتواند یاد بگیرد نمایشی برای ورودی در لایه‌های میانی ارائه دهد. با شبکه‌های دنباله‌به‌دنباله [37] پیاده‌سازی خلاصه‌سازهای چکیده بیشتر از گذشته امکان‌پذیر شده است [38] و اکثر تحقیقات با استفاده از این شبکه‌ها در حال انجام است [39].

مدل‌های شبکه عصبی [40] با ساختار رمزنگار-رمزگشا به همراه مکانیزم توجه[2] علاوه بر محبوبیت استفاده در ماشین‌های ترجمه [10]، توانسته‌اند مقدار ROUGE قابل‌توجهی را برای خلاصه‌سازی چکیده‌ای کسب کنند [41]. در ادامه به بررسی مراجعی خواهیم پرداخت که از این معماری استفاده کرده‌اند.

در مرجع [10] با استفاده از معماری نشان داده شده در شکل ۲–۹ خلاصه‌ساز چکیده‌ای پیاده‌سازی شده است که بر پایه معماری رمزنگار-رمزگشا و مکانیزم توجه پیاده‌سازی شده. همانطور که در شکل ۲–۹ مشاهده می‌شود، ورودی رمزگذار تنها از بردار جاسازی (امبدینگ) کلمات متن ورودی استفاده نشده است بلکه برای اینکه در خلاصه‌ی خروجی به مفاهیم و کلمات کلیدی دقت بیشتری شود از دیگر ویژگی‌های پردازش متنی مانند نقش‌واژه‌ها، TF و IDF استفاده شده است.

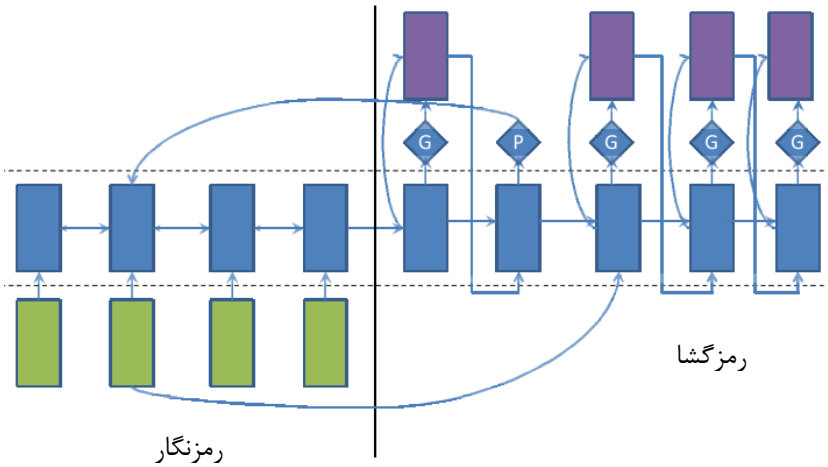
[1] Parser
[2] Atention



شکل ۲–۹ معماری رمزگذار–رمزگشا، پیشنهاد داده شده در مرجع [10] برای خلاصه‌سازی چکیده‌ای که از مکانیزم توجه بهره گرفته شده.

از دیگر خلاقیت‌های مرجع [10] استفاده از سازوکاری است که مشکل عدم وجود کلماتی که در لغت‌نامه‌ی مدل وجود (OOV[1]) ندارد را حل می‌کند. از آنجایی که ممکن است کلمات کلیدی در متن ورودی، اسم‌هایی باشد که در لغت‌نامه وجود نداشته باشند، دراینصورت در خروجی خلاصه‌ساز در صورت نبود سازوکاری برای این مساله، از این کلمات مهم چشم‌پوشی می‌شود. برای حل این مشکل، مقاله مکانیزم کلید تولید-نشانه‌گر[۲] را پیشنهاد داده است که در شکل ۲–۱۰ قابل مشاهده است.

[1] Out of vocabulary
[2] Pointer-generator switch



شکل ۲–۱۰ مکانیزم کلید تولید-نشانه‌گر پیشنهاد شده در مرجع

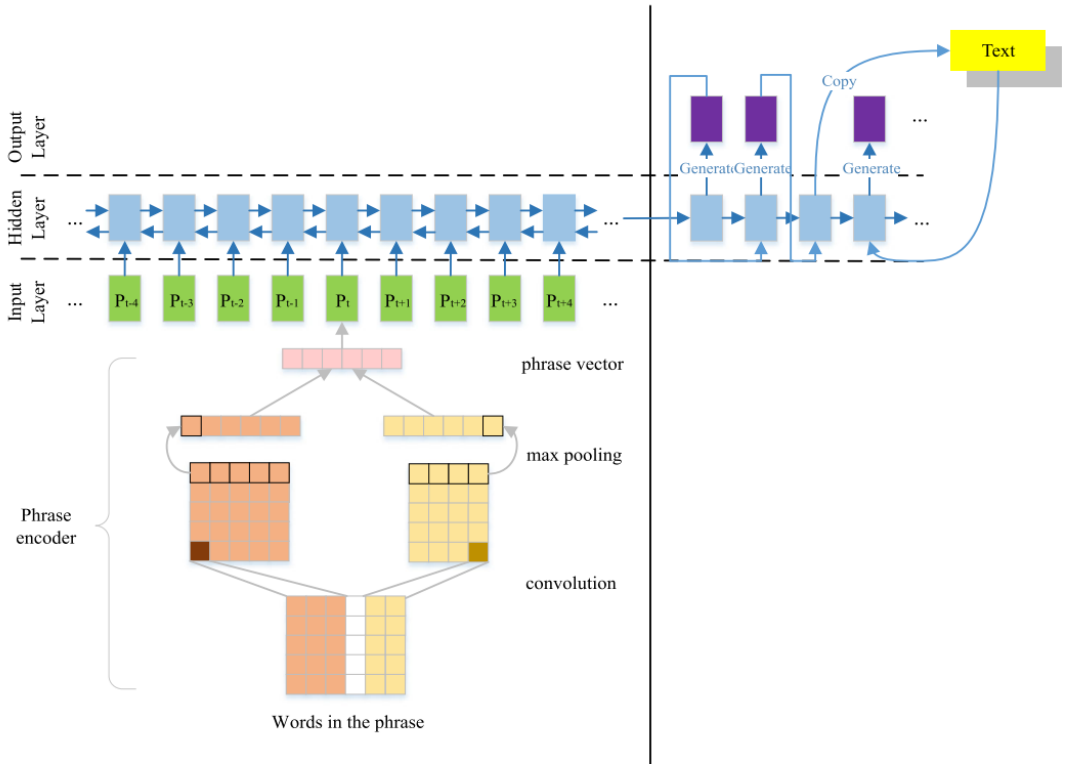
طریقه عملکرد کلید تولید-نشانه‌گر به این شکل است که یک کلید که در شکل ۲–۱۰ با علامت G و P آموزش داده می‌شود؛ این کلید در صورت که نیاز باشد از G به P تغییر کند؛ در این حالت رمزگشا بجای تولید یک کلمه در دایره واژگان داخل لغت‌نامه به صورت مستقیم از ورودی جاسازی کلمه مورد نظر را برمی‌دارد و در خروجی قرار می‌دهد. این کلید با استفاده از تابع فعال‌ساز سیگموید[1] پیاده‌سازی شده است. از مشکلات مهم مشاهده شده در مرجع [10] می‌توان به تکرار عبارات یا کلمات دراشاره کرد [41]. در مرجع [41] با تغییر مکانیزم توجه، توانسته این مشکل را کمرنگ‌تر بکند.

در مرجع [42] با استفاده از یک معماری مبتنی بر LSTM و TCN خلاصه‌ساز چکیده‌ای پیاده‌سازی شده است. در این حالت، به جای اینکه دنباله کلمات را به عنوان ورودی مدل در نظر بگیرد از دنباله[2] عبارات[3] به عنوان ورودی استفاده می‌شود. معماری مدل پیشنهاد شده در شکل ۲–۱۱ آورده شده است.

همان‌طور که در شکل ۲–۹ مشاهده می‌شود. در قسمت رمزگشا، ورودی به صورت عبارات از قبل جدا شده به‌صورت دنباله‌ای به TCN ورودی داده می‌شود تا نمایشی از آن بردار ارائه شود. در مرحله بعد، در

[1] Sigmoid
[2] Sequence
[3] Phrase

قسمت رمزگشا، نمایش ارائه شده توسط مرحله قبل به‌صورت دنباله به دنباله به یک شبکه LSTM داده می‌شود. لازم به ذکر است اول و انتهای متن ورودی با یک کاراکتر خاص مشخص می‌شود. بعد از گرفتن کاراکتری که نشان‌دهنده انتهای مدل است، مدل یک بردار که نماینده‌ای از مراحل انجام شده در رمزگشا است برای رمزنگار ارائه می‌دهد. در ادامه رمزنگار وظیفه دارد باتوجه‌به اطلاعات که از رمزنگار می‌گیرد، خلاصه‌ای ارائه دهد که مفاهیم اصلی متن مرجع را در برداشته باشد. لازم به ذکر است برای این کار از مکانیزم توجه هم استفاده شده است؛ در این مکانیزم یاد گرفته می‌شود که به بخش‌های خاصی از متن ورودی باتوجه‌به خروجی توجه بیشتری شود تا خروجی بهتری در آن لحظه تولید شود. مدل ارائه شده با معیار ROUGE-1 توانسته عدد ۳۹.۴ را کسب کند.



شکل ۲–۱۱ معماری پیشنهاد شده در مرجع [42] که از دو بخش رمزگزار (سمت چپ) و رمزگشا (سمت چپ) تشکیل شده است.

استفاده از شبکه‌های عصبی عمیق بازگشتی (RNN) و معماری رمزنگار رمزگشا یکی از راه‌حل‌های

پیاده‌سازی خلاصه‌سازهای چکیده‌ای است [4]. از معایب این نوع شبکه‌ها می‌توان به موارد زیر اشاره کرد:

- تکرار کلمه‌ها و عبارت ممکن است در خروجی ظاهر شود.
- کلماتی که در لغت‌نامه[1] (منبع واژگانی) وجود ندارند یا محدود استفاده‌شده‌اند می‌تواند مشکل‌زا باشند.
- نیاز به مجموعه دادگان عظیم برای آموزش این مدل‌ها و همچنین نیاز به منابع محاسباتی بالا می‌باشد.

با ظهور شبکه‌های از پیش‌آموزش[2] داده شده برای اعمال پردازش زبان‌های طبیعی، تحقیقات بر روی خلاصه‌سازهای چکیده‌ای به این جهت متمایل می‌شود. این مدل‌های زبانی برای اعمالی مانند پیش‌بینی کلمهٔ بعدی یا جای‌گذاری جای خالی با کلمه مناسب آموزش داده شده‌اند [43]. همچنین با تنظیم این مدل‌ها به صورت باناظر، می‌توان اعمال پایین‌دستی[3] مانند خلاصه‌سازی چکیده‌ای و مترجم خودکار را پیاده‌سازی کرد. از آنجایی که این مدل‌ها از قبل با متون عظیم آموزش داده شده‌اند؛ در نتیجه برای اعمالی پایین‌دستی مانند خلاصه‌سازی، با دادگانی محدودتری می‌توان مدل را تنظیم کرد و این یک نکته مثبت برای استفاده از این مدل‌ها است [12]. مدل زبانی T5 که در مرجع [44] معرفی شده است مناسب برای خلاصه‌سازی زبان انگلیسی است و منابعی از آن برای اینکار استفاده کرده‌اند؛ به عنوان مثال در مرجع [45] با استفاده از مدل T5 برای عمل خلاصه‌ساز چکیده‌ای با معیار ROUGE-1، بر روی دادگان CNN-DailyMail [46] عدد ۳۹ کسب شده است.

همچنین در مرجع [43] ضمن معرفی یک مدل زبانی پیش‌آموزش داده‌شده به نام PEGASUS، و تنظیم آن خلاصه‌ساز چکیده‌ای پیاده‌سازی شده است. روش آموزش این مدل به این صورت است که به جای جایگذاری کلمات حذف شده، مدل با جایگذاری جملات حذف شده آموزش می‌بیند. با توجه به اینکه

[1] Vocabulary
[2] Pre-trained
[3] Down-stream task

این‌کار مشابه خلاصه سازی استخراجی است؛ نویسندگان این مدل را مناسب برای اعمال پایین‌دستی مرتبط با خلاصه‌سازی می‌دانند. در این کار برای عمل خلاصه‌سازی مقدار ROUGH-1 ۴۲.۷ با دادگان CNN-DailyMail کسب شده است.

در مرجع [2] یک مجموعه دادگان مناسب برای پیاده‌سازی عنوان‌ساز ارائه شده است. این دادگان با خزش سایت‌های منتشرکننده مقالات علمی دادگان را جمع‌آوری کرده است. این مجموعه شامل چکیده، عنوان چکیده و همچنین کلیدواژه مقالات می‌باشند. در این کار با استفاده از یک مدل انتقالی بر روی دادگان پیشنهاد شده دقت ۳۷.۳ با معیار ROUGH1 بدست آمده است. در جدول ۲–۱ زیر دقت مدل‌های مختلف قابل مشاهده است. در این جدول مشاهده می‌شود که سه روش اول که از نوع استخراجی هستند از دو روش دیگر که چکیده‌ای هستند دقت کمتری را با معیار ROUGH-1 بدست آورده است.

جدول ۲–۱ ارزیابی دادگان پیشنهاد شده با مدل‌های مختلف درمرجع [2]

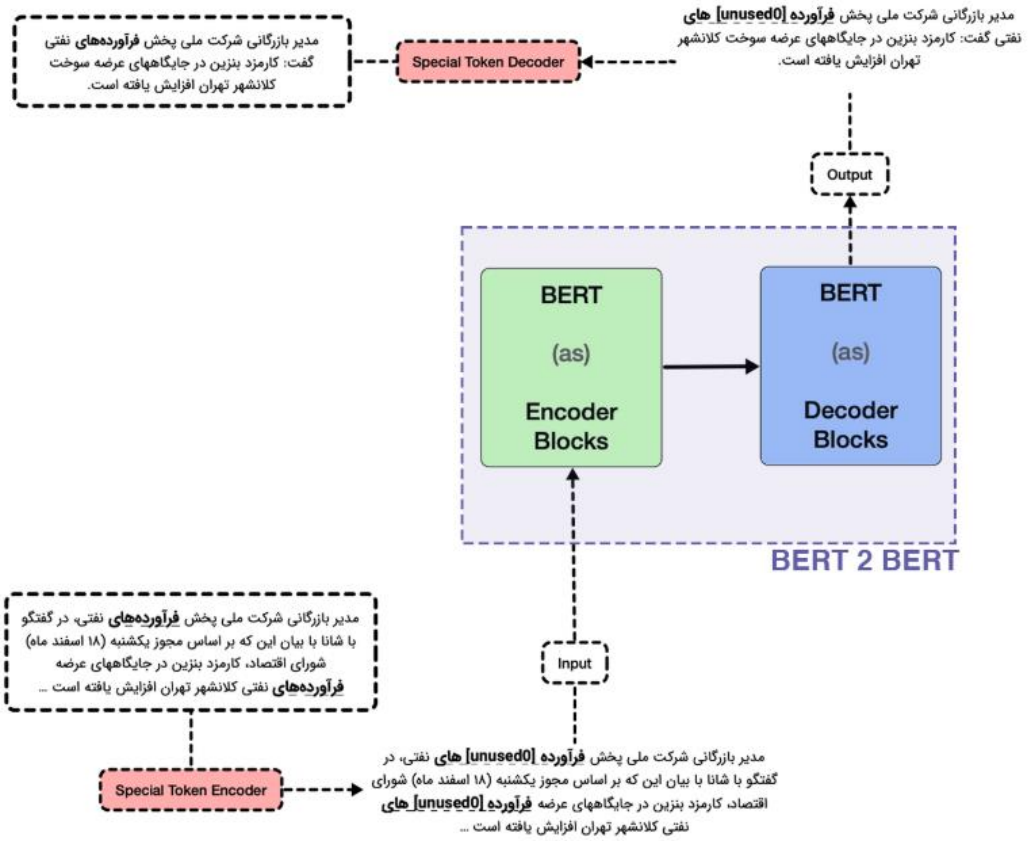
Method	R1	R2	RL
Random-1	22.67	8.02	18.44
Lead-1	33.38	16.8	28.44
LexRank	29.4	12.83	24.03
PointCov	36.12	18.88	30.21
Transformer	37.27	19.12	30.78

برای زبان فارسی که یک زبان منبع محدود به‌حساب می‌آید [47] تا بحال خلاصه‌ساز چیکده‌ای قابل توجه ارائه نشده بود تا اینکه با ظهور شبکه‌های از پیش‌آموزش داده شده، مثل mT5 [48] نیاز کثرت دادگان کمتر شد و با حجم محدودتری از دادگان می‌توان مدل را برای اعمال پایین‌دستی مانند خلاصه‌سازی تنظیم کرد [48].

در مرجع [49] برای زبان فارسی یک خلاصه‌ساز چکیده ارائه شده است. مجموعه دادگان مورد نیاز برای پیاده‌سازی، توسط محققان این مقاله جمع‌آوری شده است. در این کار علاوه بر معرفی مجموعه دادگان مناسب برای خلاصه‌سازی، با استفاده از دو مدل زبانی انتقالی mT5 [49] و BERT [14] مدل خلاصه‌ساز

چکیده‌ای فارسی ارائه شده است. در این مرجع ازParsBERT [50] که برپایه‌ی BERT می‌باشد، به عنوان رمزنگار و از یک شبکه BERT به عنوان رمزگشا استفاده کرده‌اند. مدل زبانی BERT تنها می‌تواند به عنوان رمزگشا یا رمزنگار استفاده شود ولی می‌توان هر دوی این ساختمان را یکی به عنوان رمزنگار و دیگری را به عنوان رمزگشا، به دنبال هم برای پیاده‌سازی معماری رمزنگار رمزگشا استفاده کرد؛ به این معماری Bert2Bert هم می‌گویند. لازم به ذکر است مدل ParsBERT همان مدل BERT است که با پیکرگان فارسی به صورت خودناظر تنظیم شده است و مناسب زبان فارسی می‌باشد [50]. در شکل ۲–۱۲ نمایی از معماری ارائه شده بر پایه‌ی BERT آورده شده است. همچنین در این شکل مشاهده می‌شود که در ورودی و خروجی، یک رمزنگار و رمزگشا برای در نظر گرفتن کاراکتر نیم‌فاصله (U+200C) که مختص فارسی است پیشنهاد شده است. در این کار قبل از ورودی متن فارسی پردازش می‌شود و به جای نیم‌فاصله یک قطعه دیگر جایگذاری می‌شود. همچنین در خروجی یک رمزگشا تعبیه شده که قطعه‌ی مختص به نیم‌فاصله را به کاراکتر نیم فاصله برگرداند.

همچنین در این کار به‌جز مدل پایه BERT از یک مدل mT5 هم استفاده شده. این مدل چندزبانه است و بر روی معماری T5 با مجموعه پیکره‌گان چندزبانه آموزش‌داده‌شده است. در این مدل زبان فارسی هم پیشنهاد داده شده است و همانند T5 توسط گوگل ارائه شده است.



شکل ۲–۱۲ معماری پیشنهاد شده در مرجع [49] **با استفاده ازمدل زبانی** BERT **و معماری رمزنگار رمزگشا.**

همچنین در مرجع [49] با تکنیک خزش شبکه وب در مراجع فارسی، توانسته است دادگان مناسب برای عمل خلاصه‌سازی فارسی چکیده‌ای جمع‌آوری کند. این دادگان از سایت‌های خبرگذاری فارسی جمع‌آوری شده است که داری جفت متن خبر و چکیده‌ی خبر می‌باشند. در شکل نمودار فروانی موضوعات مختلف آورده شده است. همچنین در جدول دقت مدل‌ها که با معیار ROUGH سنجیده شده است قابل مشاهده است.

جدول ۲–۲ ارزیابی مدل‌های Bert2Bert **و** mT5 **در مرجع** [49] **با معیار** ROUGE **بر روی دادگان پیشنهادی**

	ROUGE		
Model	R-1	R2	R-L
mT5	42.25	24.36	35.94
Bert2Bert	44.01	25.07	37.76

در مرجع [51] برای پیاده‌سازی عنوان‌ساز فارسی یه مدل رمزنگار رمزگشا پیشنهاد شده که در رمزگشا از مکانیزم توجه برای درک بیشتر وابستگی‌های خروجی با کلمه‌ها ورودی استفاده شده است. در این مرجع برای پیاده‌سازی مدل رمزنگار رمزگشای با مکانیزم توجه، از ماژول OpenNMT [52] پایتورچ استفاده شده است. همچنین برای آموزش این مدل ۲۷۰ هزار جفت نمونه داده‌ی، چیکده و عنوان جمع‌آوری شده است. این داده ها از مقالات علمی و پایان‌نامه‌های دانشگاهی جمع‌آوری شده است. در جدول ۲–۳ دقت‌های بدست آمده برای مدل عنوان‌ساز ارائه شده با معیار ROUGE ارائه شده. از معایب این رویکرد می‌توان به کند بودن مدل‌ها و نیاز به دادگان بیشتر اشاره کرد.

جدول ۲–۳ ارزیابی مدل ارائه شده با داده‌های پیشنهادی در مرجع [51]		
ROUGE-1	**ROUGE-2**	**ROUGE-L**
۴۲.۶	۲۵.۱	۳۹.۰

۲-۴ جمع‌بندی

در این قسمت دو رویکرد خلاصه‌سازهای استخراجی و چکیده‌ای موردبحث قرار گرفت. کارهای مهم در هر یک و همچنین برای زبان فارسی، بررسی شد. همچنین مراجعی را معرفی کردیم که دادگانی برای خلاصه‌سازی و عنوان‌سازی ارائه کرده بودند. با مرور کارهای گذشته، برای ارائه‌ی یک عنوان‌ساز فارسی که بتواند دقت خوبی داشته باشد، سعی کردیم دادگانی مناسب برای اینکار جمع‌آوری کنیم و از بروزترین مدل‌ها و تکنیک‌ها برای پیاده‌سازی عنوان‌ساز فارسی استفاده کنیم. هدف تولید عنوان‌های با کیفیت است که بتواند به خوبی برای عنوان متن مرجع انتخاب شود تا در کاربردهایی مانند پیشنهاد عنوان متن ایمیل در نرم‌افزارهای میزبان ایمیل برای زبان فارسی استفاده شود.

در جدول ۲–۴ مقایسه از کارهای انجام شده‌ی در خلاصه‌سازی خودکار با رویکردهای استخراجی و چکیده‌ای با زبان‌ها فارسی و انگلیسی ارائه شده است. با توجه به اینکه در این مراجع برای ارزیابی مدل

ارائه شده از دادگان و معیارهای مختلفی استفاده شده است؛ در کنار بهترین دقت ارائه شده در مرجع، نوع معیار و دادگانی که استفاده شده است هم آورده شده.

جدول ۲–۴ شرح خلاصه‌ای از فعالیت‌های کارهای پیشین

زبان	مرجع	توضیحات	دقت	معیار	دادگان
انگلیسی	[20]	[استخراجی] در این مرجع فرض شده تکرار کلمات اهمیت آن کلمه و جمله‌ای که در آن هست را افزایش می‌دهد.	۴۳	Precision 1-gram	DUC-2002
	[23]	[استخراجی] نمایش متون با استفاده از استخراج اطلاعات آماری به عنوان ویژگی، انجام شده است.	۳۶	ROUGE-1	DUC-2007
	[24]	[استخراجی] نمایش متون با استفاده از نظریه‌گراف‌ها صورت گرفته و از دیتابیس واژگانی WordNet برای استخراج روابط معنایی استفاده شده است .	۴۰	Precision 1-gram	DUC-2002
	[22]	[استخراجی] جملات با ویژگی‌های آماری نمایش‌داده می‌شوند. همچنین مدلی آموزش داده می‌شود تا جملات مهم را رتبه‌بندی کند.	۴۷	ROUGE-1	DUC-2002
	[36]	[چکیده‌ای] در این روش ساختار نحوی متن ورودی با نظریه گراف‌ها استخراج و نمایش داده می‌شوند. جملات با معانی نزدیک خوشه‌بندی می‌شوند. در نهایت با استفاده از مدل‌های تولید زبان برای هر نماینده خوشه‌ها یک جمله برای خلاصه نهایی تولید می‌شود	۳۲	ROUGE-1	دادگان پیشنهادی
	[6]	[چکیده‌ای] از درخت وابستگی برای نمایش جملات استفاده می‌شود. جملات با معانی نزدیک خوشه‌بندی می‌شوند. در نهایت با استفاده از مدل‌های تولید زبان برای هر نماینده خوشه‌ها یک جمله برای خلاصه نهایی تولید می‌شود.	۴۳	ROUGE-1	DUC-2007
	[5]	[چکیده‌ای] خلاصه‌ساز چند متنه ارائه شده است. از ساختار نهادگزاره برای نمایش جملات استفاده شده است.	۸۵	Pyramid	DUC-2002
	[10]	[چکیده‌ای] خلاصه‌ساز چکیده‌ای پیاده‌سازی شده است که بر پایه معماری رمزنگار-رمزگشا و مکانیزم توجه است.	۳۵	ROUGE-1	CNN-DailyMail
	[42]	[چکیده‌ای] بر پایه معماری رمزنگار-رمزگشا و استفاده از ساختارهای LSTM و TCN خلاصه‌ساز چکیده‌ای پیاده‌سازی	۳۹	ROUGE-1	CNN-DailyMail

		شده است.			
	[45]	[چکیده‌ای] از مدل انتقای پیش‌آموزش داده شده T5 برای پیاده‌سازی خلاصه‌ساز استفاده شده است.	۳۹	ROUGE-1	CNN-DailyMail
	[43]	[چکیده‌ای] مدل زبانی انتقالی مناسب برای خلاصه‌سازی پیاده‌سازی شده است.	۴۲	ROUGE-1	CNN-DailyMail
فارسی	[26]	[استخراجی] نمایش متون با استفاده از نظریه‌گراف‌ها صورت گرفته و از دیتابیس واژگانی FarsNet برای استخراج روابط معنایی استفاده شده است.	۶۵	Precision of common sentences	دادگان پیشنهادی
	[32]	[استخراجی] برای نمایش متون از جاسازی استفاده می‌شود. سپس با استفاده از شبکه‌ای کلمات موجودیت نام گذاری شده (NER) شناسایی می‌شود و بر اساس تعداد این کلمات جملات رنک‌بندی می‌شوند.	۴۶	ROUGE-1	دادگان پیشنهادی
	[49]	[چکیده‌ای] با استفاده از مدل‌های زبانی انتقالی mT5 و BERT2BERT خلاصه‌ساز چکیده‌ای ارائه شده است.	۴۴	ROUGE-1	دادگان پیشنهادی
	[51]	[چکیده‌ای] با استفاده از یک شبکه رمزنگار رمزگشا بر پایه شبکه‌های بازگشتی LSTM عنوان‌ساز فارسی ارائه شده است.	۴۲	ROUGE-1	دادگان پیشنهادی

فصل ۳ ادبیات موضوع

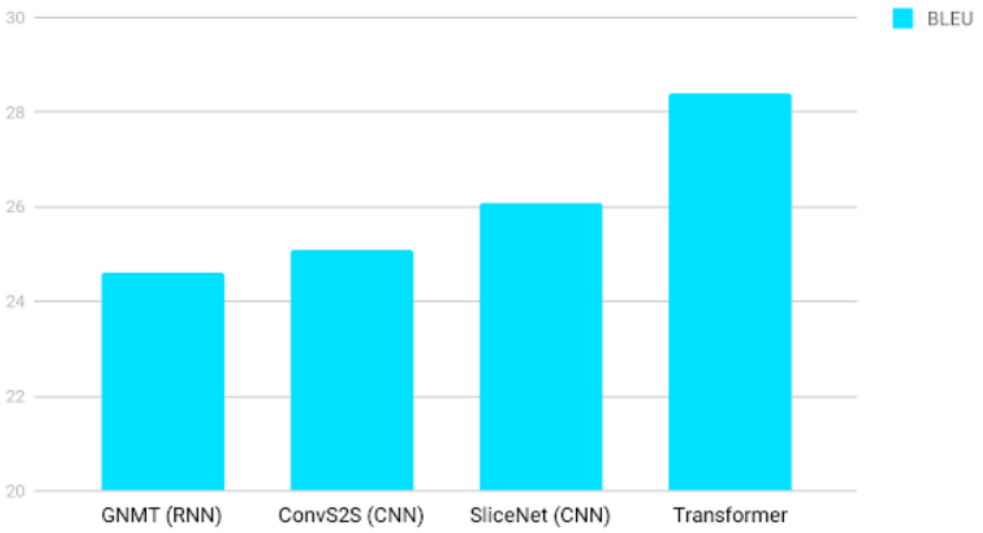
۳-۱ مقدمه

تا قبل از انتشار مرجع [11] "توجه همان چیزی است که می‌خواهید[1]" توسط شرکت گوگل، شبکه‌هایی بازگشتی مانند LSTM[2] و GRU[3] بهترین رویکرد برای درک زبان و پیاده‌سازی مدل‌هایی مانند، مترجم خودکار، تولید متن و سیستم‌های سوال و پاسخ به حساب می‌آمدند. تمام تلاش‌ها برای بهبود این شبکه‌ها انجام می‌شد تا دقت خروجی در این ساختارها افزایش پیدا کند. این دسته از شبکه‌های بازگشتی (RNNs, LSTMs) به طور معمول موقعیت ورودی‌ها و خروجی‌ها را در نظر نمی‌گرفتند؛ همچنین با توجه به اثر محو شدن گرادیان[4] در شبکه‌های بازگشتی، اطلاعات دورتر کمرنگ‌تر می‌شود و عملا دستیابی به آنها سخت‌تر می‌شود. در مرجع [11] یک معماری شبکه عصبی به نام مدل انتقالی[5] معرفی شد که برای درک زبان‌ها طبیعی به خوبی عمل می‌کند. در شکل ۳–۱دقت مدل‌های ترجمه خودکار، پیاده‌سازی شده با معماری مدل‌های انتقالی و دیگر معماری‌ها به نمایش گذاشته شده است؛ همانطور که مشاهده می‌شود معماری انتقالی توانسته است با اختلاف قابل توجهی نسبت به مدل‌های دیگر، دقت خوبی را با معیار BLEU کسب کند (هر چه عدد معیار BLUE بیشتر باشد مدل بهتر است).

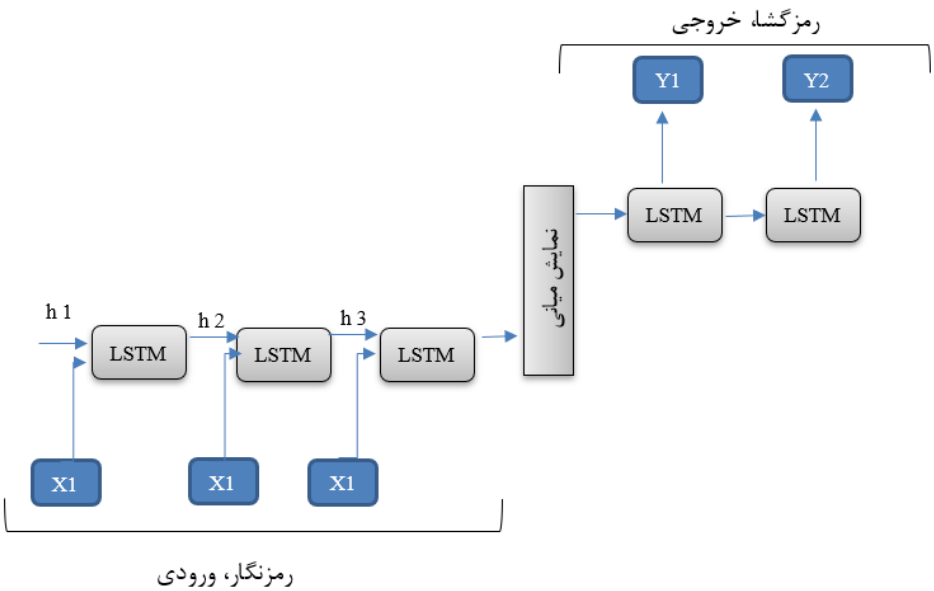
شبکه‌های بازگشتی مانند RNN و LSTM ماهیت دنباله‌ای دارند؛ در کاربردهای پردازش متن کلمات متون به صورت دنباله‌به‌دنباله به ترتیب به مدل داده می‌شود؛ در این معماری مدل باید سعی کند روابط بین کلمات گذشته و آینده (در صورت دو مسیره بودن) را استخراج کند؛ تشخیص روابط با توجه به مساله محو شدن گرادیان سخت‌تر می‌شود. هرچه دنباله کلمات طولانی‌تر باشد مدل سخت‌تر می‌توانند روابط بین کلمات را مدل کند. همچنین به دلیل ماهیت دنباله‌ای، این مدل‌ها سازگار با پردازش موازی به‌روز توسط دستگاه‌های GPU و TPU نمی‌باشند. در شکل ۲–۳یه نمونه معماری از شبکه‌های رمزگذار رمزگشا بر

[1] Attention is all you need
[2] Long Short-Term memory
[3] Gated Recurrent Units
[4] Vanishing gradient
[5] Transformer

پایه شبکه‌های بازگشتی آورده شده است. این شبکه‌ها تا قبل از شبکه های انتقالی برای پیاده‌سازی خلاصه‌سازها و ماشین‌های ترجمه خودکار مجبوبیت داشتند.



شکل ۳–۱ مقایسه دقت خلاصه‌سازهای خودکار با مدل انتقالی و دیگر مدل‌های رایج مانند RNN در مرجع [11]



شکل ۳–۲ یک نمونه از معماری رمزنگار–رمزگشا برپایه شبکه‌های بازگشتی

برخلاف شبکه‌های بازگشتی، شبکه‌های مبتنی بر معماری انتقالی، تعداد کلمات مشخصی را به صورت یکجا

دریافت می‌کند و برای آن‌ها جاسازی ارائه می‌دهد. همچنین با استفاده از مکانیزم توجه روابط بین کلمات ورودی بدون توجه به دوری و نزدیک آنها در جمله، یاد گرفته می‌شود؛ و نهایت در رمزگشا با استفاده از نمایش ارائه در مرحله قبل (رمزنگار) کلمات به ترتیب تولید می‌شوند. همچنین در این مرحله، برای پیشبینی هر کلمه به کلمات ارائه شده در ورودی و همچنین کلمات پیش‌بینی شده قبلی، توجه می‌شود (مکانیزم توجه).

در بخش بعدی، به معرفی مدل‌های انتقالی خواهیم پرداخت که از مکانیزم توجه بهره گرفته‌اند. همچنین به معرفی معیار ROUGE که برای ارزیابی خلاصه‌های ماشینی کاربرد دارند خواهیم پرداخت.

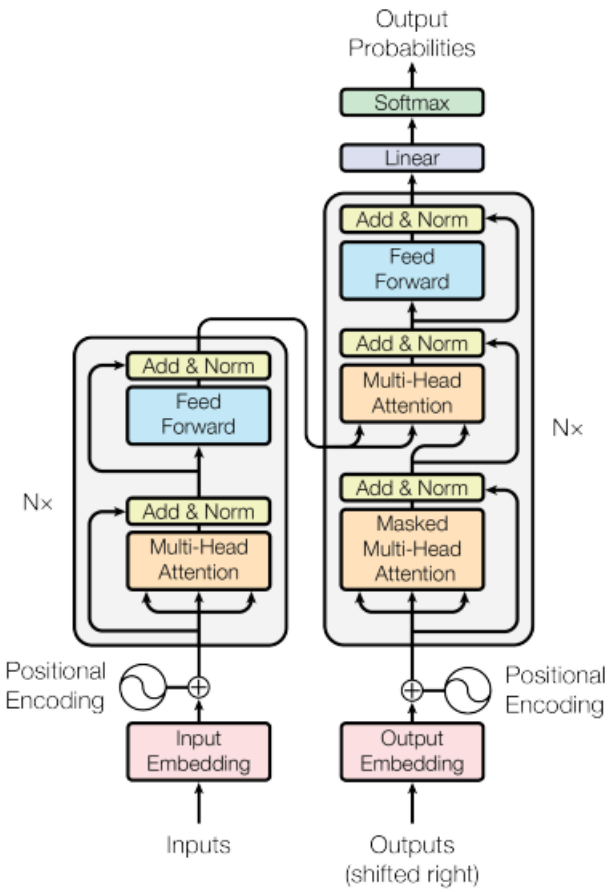
۳-۲ مدل‌های انتقالی

مدل‌های رمزنگار رمزگشا بر پایه شبکه‌های بازگشتی، از آن جهت که داده‌ها به صورت دنباله‌ای به مدل داده می‌شود، سرعت پایین‌تری دارند. همچنین به صورت واقعی چند جهته نیستند که بر خلاف ماهیت متون به حساب می‌آید [11]. برای رفع این مشکلات یک معماری در مقاله [11] پیشنهاد شده که سریع‌تر از شبکه‌های پایه بازگشتی هستند به این دلیل که ماهیت دنباله‌ای ندارند؛ همچنین عملکردی چندجهته دارند از آنجایی که بدون در نظر گرفتن موقعیت کلمات، روابط بین آن‌ها استخراج می‌شود. این معماری در شکل ۳–۳ آورده شده است. اجزای این مدل از قرار زیر است:

- رمزنگار: از روی هم قرار گذاشتن N لایه تشکیل می‌شود. در مرجع [11]، $N = ۶$ در نظر گرفته شده است. هر لایه از زیر لایه تشکیل شده است. در این قسمت علاوه برای زیرلایه‌ی تماما متصل[1] و نرمال‌کننده از لایه‌ی خودتوجه[2] هم استفاده شده تا روابط بین کلمات ورودی‌ها استخراج شود. در نهایت خروجی این مدل به صورت جاسازی شده برای استنتاج به رمزگشا داده می‌شود.

[1] Fully connected
[2] Self attention

- رمزگشا: مانند بخش رمزنگار از روی هم قرار دادن N لایه تشکیل شده است. که هر لایه علاوه بر دولایه‌ای که در رمزگشا هم وجود دارد یک لایه توجه اضافی برای بررسی ارتباطات بین خروجی و ورودی‌ها در لایه رمزنگار قرار داده شده است.



شکل ۳–۳ معماری مدل انتقالی مرجع [11]

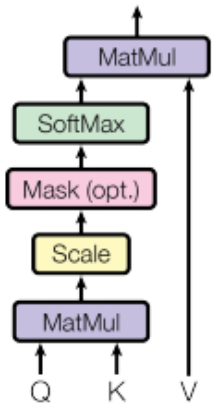
همانطور که گفته شد در مدل‌ها انتقالی با استفاده از این مکانیزم توجه، مسیرها و دسترسی ها به اطلاعات کوتاه‌تر و مستقیم‌تر می‌شود و بدون هیچ محدودیتی مدل می‌تواند در وضعیت فعلی به اطلاعات خروجی‌های قبلی و ورودی‌های رمزگشا، برای تولید خروجی جدید دست پیدا کند. با کوتاه شدن بازیابی اطلاعات پیچیدگی‌های زمانی کاهش پیدا کرده است و به صورت قابل توجهی زمان آموزش مدل کاهش پیدا کرده است. همچنین علاوه بر کاهش پیچیدگی تعداد عملیات‌های دنباله‌ای کاهش پیدا می‌کند (جدول ۳–۱).

جدول ۳–۱ مقایسه‌ی مدل دارای مکانیزم توجه با مدل‌های سنتی بازگشتی مرجع [11]

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

۳-۲-۱ مکانیزم توجه

یک تابع توجه را می توان به صورت نگاشت یک بردار کوئری (Q) و مجموعه ای از جفت های کلید-مقدار (key-value) به یک خروجی توصیف کرد، جایی که Q، K، V، و خروجی ها همه بردار هستند. خروجی به عنوان یک جمع وزنی از مقادیر محاسبه می‌شود، جایی که وزن اختصاص داده شده به هر مقدار توسط یک تابع تولید می‌شود (شکل ۳–۴).



شکل ۳–۴ تابع توجه مرجع [۱٤]

در این تابع ورودی‌ها متشکل از Q ,V,K می‌باشند رابطه ۳–۱ خروجی به عنوان یک جمع وزنی از مقادیر V ها محاسبه می‌شود، جایی که وزن اختصاص داده شده به هر مقدار توسط ضرب داخلی Q و K محاسبه می‌شود؛ این ضرب توسط $\sqrt{d_k}$ اسکیل شده است (در اینجا d_k ابعاد بردار K می‌باشد) . همچنین

ضرب داخلی به یک تابع سافت‌مکس[1] ورودی داده شده است و در خروجی در V ضرب می‌شود. قابل ذکر است که همه‌ی بردارهای V,K,Q دارای ابعاد یکسانی هستند.

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

رابطه ۳–۱ تابع توجه

به صورت شهودی‌تر مقدار Q از مرحله‌ی قبل می‌آید و خبر از نیاز به یک وابستگی به حالت‌های قبلی می‌دهد؛ همچنین مقدار V بردارهایی هستند که خاصیت حالت‌های قبلی را نشان می‌دهند و همچنین بردارهای K در واقع نماینده‌ای از هر ستون V هستند. در اینجا با استفاده از ضرب داخلی Q و K شباهت بین بردارهای Q و V سنجیده می‌شود و در ادامه با استفاده از تابع سافت‌مکس آن تابعی که شباهت بیشتری داشته باشد پرنگ‌تر می‌شود. در نهایت مقدار بدست آمده در مرحله‌ی قبل وزن‌های هر سطر ماتریس V را نشان می‌دهد که هر یک میزان توجه بیشتر یا کمتر را برای استنتاج در مرحله‌ی فعلی مشخص می‌کند.

۳-۳ مدل انتقالی T5

در مرجع [12] مدل زبانی به نام T5[2] معرفی شده است که از معماری انتقالی (شکل ۳–۳) رمزنگار و رمزگشا که در بخش قبل توضیح داده شده، استفاده شده است. این مدل در مرحله آموزش با استفاده از پیکرگان عظیم C4[3] به صورت خودناظر آموزش داده شده. یک از روش‌های آموزش این مدل به این صورت که کلماتی به صورت رندم از متن حذف می‌شود و مدل وظیفه دارد در این قسمت به صورت با نظارت برای تشخیص پیش‌بینی کلمه‌ی حذف شده بهینه شود. به این دلیل این فرآیند را خودناظر می‌نامند که مدل از

[1] Softmax function

[2] Text-to-Text Transfer Transformer

[3] Colossal Clean Crawled Corpus (C4)

درون مجموعه داده آموزش برچسب و داده‌های آموزش را بیرون می‌کشد. در اینجا برچسب کلمه‌ی حذف شده و ورودی جمله یا جمله‌هایی با جای خالی هستند. این مدل عظیم نسخه base آن شامل حدود ۲۰۰ میلیون پارامتر می‌باشد و نسخه‌های large و small هم برای استفاده عموم ارائه شده است.

۳-۳-۱ معماری مدل

برای این مدل از بلاک‌های رمزنگار و رمزگشا معرفی شده در مقاله [11] (شکل ۳–۳) استفاده شده است. ۱۲ بلوک رمزگشا و ۱۲ بلوک رمزنگار بر روی هم گذاشته شده‌اند تا یک مدل رمزنگار رمزگشای واحد به نام T5 ارائه شود. در هر بلاک، دو لایه شبکه‌های تماما متصل استفاده شده است؛ هر لایه ابعادی به اندازه ۳۰۷۲ نورن به همراه تابع غیر خطی RELU می‌باشند. ماتریس‌های key و value، در مکانیزم توجه دارای ابعاد ۶۴ با ۱۲ head می‌باشند. ابعاد تمامی زیر لایه‌ها و جاسازی‌ها ۷۶۸ تایی می‌باشد. در مجموع این مدل (mT5-bass) دارای ۲۲۰ میلیون پارامتر می‌شود. برای آموزش این مدل از روش بهینه‌سازی AdaFactor و تابع خطای cross-entropy استفاده شده است.

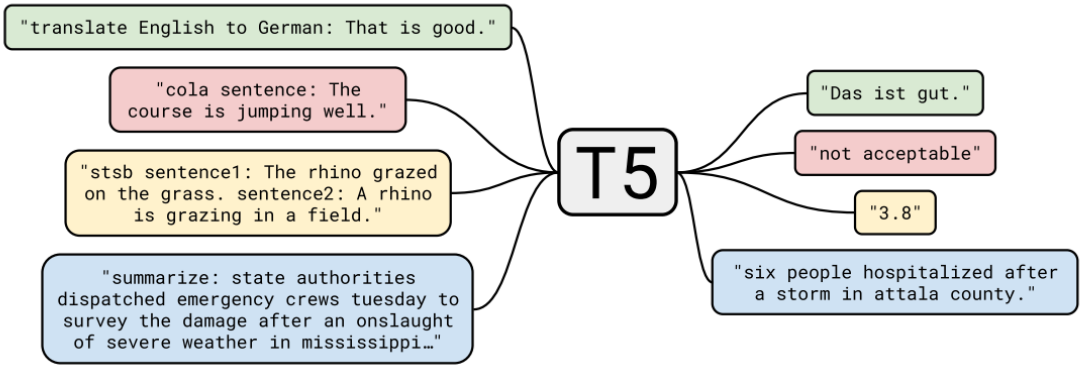
۳-۳-۲ اعمال پایین‌دستی[1]

منظور از اعمال پایین‌دستی در رابطه با مدل‌های انتقالی این است که مدل قادر به انجام این اعمال در کنار اعمالی است که مدل از ابتدا برای آن آموزش دیده است (یکی از این اعمال پایه جایگذاری جای خالی با کلمه مناسب در جمله می‌باشد.) از این مدل برای اعمال پایین‌دستی مختلفی مانند، مترجم خودکار، آنالیز احساسات[2] و خلاصه‌سازی چکیده‌ای حتی بدون تنظیم می‌توان مورد استفاده قرار گیرد. محققان این کار،

[1] Downstream tasks
[2] Sentiment anlasis

مدل را با معماری متن‌به‌متن[1] طراحی و ارائه کرده‌اند. به این صورت که مدل یک مجموعه یکپارچه است که ورودی آن متن و به دنبال آن در خروجی متن تولید می‌شود. تصویری از این رویکرد در شکل ۳–۵ آورده شده است. همانطور که مشاهده می‌شود مدل پیش‌آموزش شده را می‌توان مستقیما برای اعمال پایین‌دستی استفاده کرد. به این صورت که در ابتدای ورودی به عمل پایین دستی مورد نظر اشاره می‌شود تا مدل با پردازش متن ورودی و توجه به عمل اشاره شده، خروجی متن را تولید کند. به عنوان مثال برای ترجمه متن انگلیسی به آلمانی تنها کافی است که به ابتدای ورودی متن انگلیسی رشته‌ی کلید" :Translate english to german" اضافه شود. این از آنجایی می‌آید که محققان این مدل برای اعمال پایین‌دستی مانند ترجمه انگلیسی به آلمانی و خلاصه‌ساز انگلیسی، مجددا مدل را آموزش داده‌اند طوری که به ابتدای دادگان آموزشی اعمال مورد نظر، کلید رشته مناسب چسبانده شده بوده‌اند. به همین جهت مدل قادر به شناسایی عمل مورد نظر با چسباندن کلید آن به ابتدای ورودی می‌باشد و انتظار می‌رود که در خروجی عمل مورد نظر بر روی متن انجام شود.

[1] Text-to-Text



شکل ۳–۵ معماری متن-به-متن پیشنهاد شده در مرجع . ورودی به صورت متن به اضافه عمل پایین‌دستی مورد نظر مستقیما ورودی داده می‌شود تا در خروجی متن به صورت مستقیما با توجه به عمل پایین‌دستی اشاره شده در ورودی، تولید شود.

۳-۴ معیار ارزیابی ROUGE[1]

معیار ROUGE ابتدا توسط مرجع [53] معرفی شد. این معیار برای ارزیابی خلاصه‌های تولید شده توسط ماشین در مقایسه با خلاصه‌هایی که توسط انسان نوشته‌شده‌است کاربرد دارد. به صورت کلی این روش با محاسبه همپوشانی[2] کلمات یا دسته‌ای از کلمات بین خلاصه‌ی ماشینی و خلاصه‌ی مرجع، معیاری را برای ارزیابی خلاصه‌ی هدف ارائه می‌دهد. در رابطه طریقه محاسبه معیار ROUGE-N آورده شده است. به صورت خلاصه‌ای این معیار همانند معیار Recall همپوشانی N-gram های موجود در متن مرجع و خلاصه‌ی خروجی را محاسبه می‌کند.

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

رابطه ۳–۲ محاسبه ROUGE-N

[1] Recall-Oriented Understudy for Gisting Evaluation
[2] Overlap

فصل۴ : روش پیشنهادی

۴-۱ مقدمه

در فصل اول مقدمه‌ای بر خلاصه‌ساز خودکار بیان شد و همچنین چالش‌های خلاصه‌ساز فارسی چکیده‌ای را شرح دادیم. همچنین در فصل دوم و سوم رویکردهای پیشین و ادبیات این موضوع را برای این مسئله، مورد بررسی قرار دادیم.

همان‌طور که گفته شد برای پیاده‌سازی خلاصه‌ساز فارسی (در مسئلهٔ ما به‌صورت دقیق‌تر عنوان‌ساز فارسی) از نوع چکیده‌ای، مدل‌های زبانی دنباله‌به‌دنباله[1] بهترین ابزار هستند [43]. برای آموزش این مدل‌ها از ابتدا علاوه بر نیاز به پیکرگان عظیم، منابع محاسباتی عظیم باید در دسترس باشد تا تحقق بپذیرد. به عنوان مثال مدل زبانی GPT-2 دارای 1.5 بیلیون پارامتر است و بر روی ۸ میلیون صفحه‌ی وب آموزش داده شده است. در اینصورت، تنها شرکت‌های بزرگ مانند OpenAi و گوگول برای این مدل‌های بزرگ منابع کافی در اختیار دارند تا مدل را آموزش دهند [13]. به همین خاطر، بجای آموزش از ابتدا، به ما این فرصت داده شده است تا با تنظیم این مدل‌ها برای کارها پایین دستی مانند خلاصه‌سازی، دسته‌بندی و غیره از این مدل‌ها بهره بگیریم. در همین جهت، در این کار تصمیم گرفتیم از مدل زبانی mT5 برای پیاده‌سازی عنوان‌ساز فارسی چکیده‌ای استفاده نماییم. این مدل توسط شرکت گوگل آموزش داده شده است و همان نسخه‌ی چند زبانه‌ی مدل T5 است. برای آموزش این مدل از پیکره‌گان mC4 استفاده شده که از خزش شبکه وب به دست آمده است. همچنین این پیکره‌گان ۱۰۱ زبان را در برمی‌گیرد که فارسی را هم جزئی از آنهاست. با توجه به چند زبانه بودن، ما این مدل را برای پیاده‌سازی عنوان‌ساز چکیده‌ای فارسی انتخاب کردیم تا با تنظیم آن در جهت کار خود بهره بگیریم.

برای این کار، نیاز به دادگانی مناسب احساس می‌شد؛ به طوری که علاوه بر تعداد نمونه‌های مناسب دارای تنوع موضوعی بالا باشند. در همین راستا، اقدام به جمع‌آوری دادگانی متشکل از فراداده‌های[2] ۷۰

[1] Sequence to Sequence
[2] Metadata

هزار مقالهٔ علمی فارسی از سایت‌های ژورنال‌های فارسی کردیم. نمایی از ناشرهای استفاده شده در مجموعه دادگان در شکل ۴–۱ قابل مشاهده است. برای پیاده‌سازی اینکار چکیده و عنوان متناظر هر نمونه‌ی دادگان را برای آموزش و ارزیابی مدل کنار هم گذاشتیم. همچنین در جهت بهبود خروجی مدل، راهکارهایی در قالب پس‌پردازش ارائه شد. به عنوان مثال یک مدل نویزگیری آموزش داده شد تا خروجی بهتری ارائه شود.

بعد از تنظیم مدل mT5، شبکه با استفاده از داده‌های تست و معیار ROUGH مورد ارزیابی قرار گرفت. در نهایت با معیار ذکر شده، مدل عدد ۴۵ را کسب کرده که سه واحد از آخرین خلاصه‌ساز چکیده‌ای فارسی ارائه شده بیشتر است.

در این فصل، ابتدا جمع‌آوری و تحلیل داده‌های بحث می‌شود. سپس دو مدل برای، عنوان‌ساز چکیده‌ای معرفی و شرح داده می‌شود.

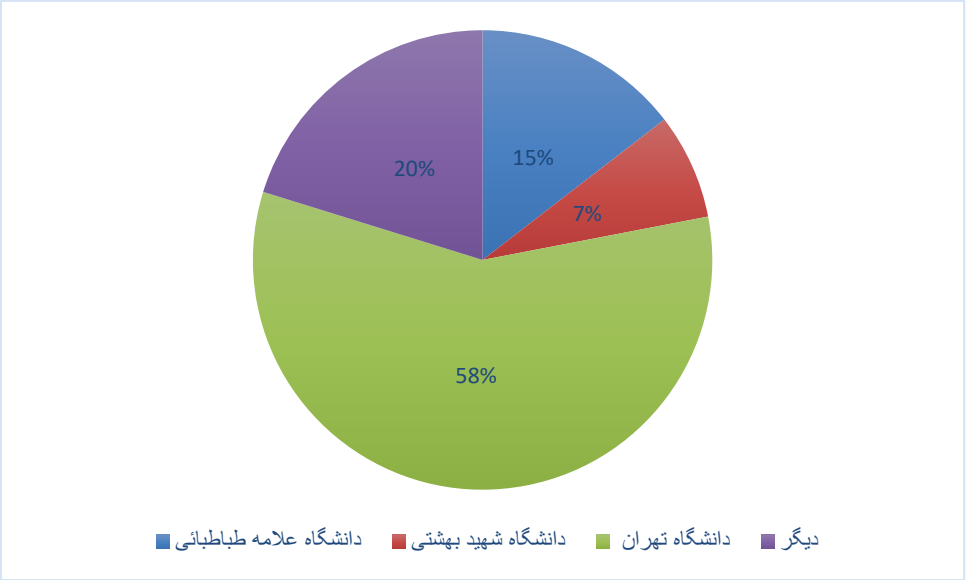
۴-۲ جمع‌آوری دادگان

همان‌طور که گفته شد مقاله‌های علمی می‌تواند داده‌های مناسبی برای پیاده‌سازی خلاصه‌سازی باشد. از آنجایی سایت‌های ژورنال‌ها اکثر اطلاعات مقالات منتشر شدهٔ خود را در دسترس عموم قرار می‌دهند؛ می‌توان با استفاده از تکنیک‌های خزش[1] فایل مقالات و اطلاعات آن‌ها را جمع‌آوری کرد. در ادامه اطلاعات حدود ۷۰ هزار مقاله از ناشرها و ژورنال‌های معتبر علمی فارسی استخراج شد. در شکل ۴–۱ سهم هر ناشر در دادگان جمع‌آوری‌شده نشان داده شده است. در جدول ۴–۱ اطلاعات اولیه از دادگان آورده شده است.

جدول ۴–۱ اطلاعات اولیه دادگان جمع‌آوری شده

تعداد نمونه‌ها	۷۳۴۶۱
تعداد ویژگی‌ها	۱۰

[1] Crawling

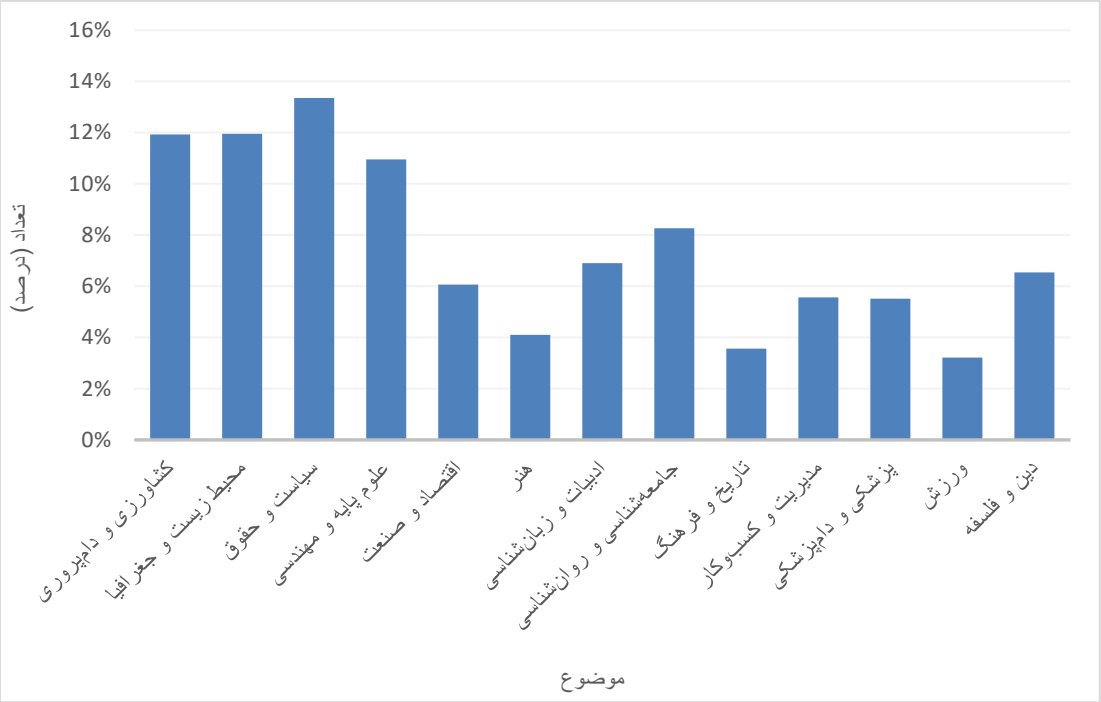


شکل ۴–۱ نمودار دایره‌ی، سهم هر ناشر از کل مقالات جمع‌آوری شده

سعی شد داده‌های جمع‌آوری شده در بازه موضوعی متنوعی باشند تا مدل آموزش‌دیده شده قادر باشد در موضوع‌های مختلف به‌خوبی عمل کند و عنوان مناسب را تولید کند. در شکل ۴–۲ تنوع موضوعی داده‌ها نمایش داده شده است.

همچنین این دادگان با استفاده از کتابخانه scrapy[1] در پایتون جمع‌آوری شده. در ابتدا فراداده‌های[2]، مقالات موجود در سایت نشریات هدف، در یک فایل json ذخیره و جمع‌آوری شد. این فایل فرمتی شبیه دیکشنری پایتون دارد ({“schema”: {schema}, “data”: {data}}) که سطرهای آن به فراداده‌ی یک مقاله اختصاص داده شده است. بدون در نظر گرفتن مقادیر نامعلوم، هر نمونه شامل چکیده، عنوان، کلیدواژه، لینک دانلود مقاله و غیره می‌شود. نمایی کلی از اطلاعات و فراداده هر نمونه در جدول ۴–۲ آورده شده است.

[1] https://github.com/scrapy/scrapy
[2] Metadate



شکل ۴–۲ تنوع موضوعی مقالات جمع‌آوری شده همانطور که ملاحظه می‌شود داده‌های دارای تنوع موضوعی گسترده می‌باشند و هر کدام از دسته‌بندی موضوعی مقداری بیش از ۱۰۰۰ نمونه را به خود اختصاص داده‌اند.

جدول ۴–۲ اطلاعات ذخیره شده برای هر مقاله

ستون	نوع متغیر	تعداد مقادیر نامعلوم	توضیح
download_url	str	۱۷۷۹۷	آدرس دانلود مقاله
file_name	str	۱۷۷۹۷	اسم فایل مقاله دانلود شده
keywords_en	str array	۷۲۷۴۷	کلیدواژه‌ها به انگلیسی
keywords_fa	str array	۷۴۵۲	کلیدواژه‌ها به فارسی
number	int	۲۰۰۳۵	شماره ژورنال منتشر شده
summary_en	str	۷۲۹۰۵	چکیده انگلیسی مقاله
summary_fa	str	۴۲۵۳	چکیده فارسی مقاله
title_en	str	۷۳۱۶۴	عنوان انگلیسی مقاله
title_fa	str	۱	عنوان فارسی مقاله
volume	int	۲۰۰۳۵	دوره ژورنال منتشر شده

۴-۳ پیش‌پردازش و بررسی آماری دادگان

۴-۳-۱ پیش‌پردازش‌های اولیه

بعد از جمع‌آوری داده‌ها با هدف پیاده‌سازی عنوان‌ساز فارسی داده‌های خام را مورد پیش‌پردازش قرار دادیم تا برای آموزش مدل آماده شود. ازآنجایی‌که قرار است به روش یادگیری با ناظر مدل را آموزش دهیم؛ جفت ورودی و خروجی مرجع را به ترتیب چکیده و عنوان مقاله در نظر می‌گیریم. به همین جهت آن نمونه‌هایی که دارای جفت اطلاعات چکیده و عنوان فارسی هستند را نگه می‌داریم و دیگر ویژگی‌ها غیر از این دو را پاک می‌کنیم. بعد از حذف مقادیر نامعلوم ۶۹۲۰۸ نمونه از کل باقی ماند.

در مرحله بعد، تمام نمونه‌ها را نرمال‌سازی می‌کنیم. یکی از نمونه‌های اصلاحات انجام‌شده در این مرحله، اصلاح نمونه‌ها از لحاظ نیم‌فاصله است. به‌عنوان‌مثال، اسم‌های جمع «فناوری ها» در واقع یک کلمه به‌حساب می‌آید به‌صورت «فناوری‌ها» اصلاح می‌شود که بین «فناوری» و «ها» کاراکتر نیم‌فاصله گذاشته می‌شود. همینطور کاراکترهای استفاده‌شده، در این مرحله یکسان می‌شوند. برای نرمال کردن متن از کتابخانه پردازش زبان فارسی هضم[1] در زبان برنامه‌نویسی پایتون استفاده شد.

بعد از مرحله حذف نمونه‌ها با مقادیر نامعلوم و نرمال‌سازی، برای اعمال یک سری پردازش‌ها و اطلاعات آماری داده‌ها را قطعه‌بندی[2] کردیم. قطعه‌بندی به این صورت انجام می‌شود که متون به واحدهایی کوچک‌تر به نام قطعه تقسیم می‌شود. در جدول ۴–۳ دو نمونه متن نرمال‌سازی‌شده به منظور قطعه‌بندی آورده شده است.

جدول ۴–۳ دو نمونه از متون پیش‌پردازش شده، قطعه بندی و اطلاح نیم‌فاصله.

متن پیش‌پردازش شده	متن خام
["اصلاح"، "نویسه‌ها"، "و"، "استفاده"، "از"، "نیم‌فاصله"، "پردازش"، "را"، "آسان"، "می‌کند"]	"اصلاح نویسه‌ها و استفاده از نیم‌فاصله پردازش را آسان می کند"

[1] https://github.com/sobhe/hazm
[2] Tokenize

["ولی", "برای", "پردازش", "،", "جدا", "بهتر", "نیست", "؟"]	"ولی برای پردازش، جدا بهتر نیست؟"

همچنین نمونه‌هایی که چکیده‌ای بین ۸۵ تا ۴۳۵ و عنوانی در بازهٔ ۳-۲۵ نداشتند را پاک کردیم. با این کار نمونه‌هایی که دارای مقادیر اشتباه هستند و نویز طلقی می‌شوند از بین داده‌ها حذف می‌شوند.

۴-۳-۲ بررسی آماری دیتاست

در ادامه بعد از انجام پیش‌پردازش و قطعه‌بندی به استخراج اطلاعات آماری دیتاست می‌پردازیم. در جدول ۴–۴ اطلاعات آماری نمونه‌ها در سطح قطعه[1] جمع‌آوری شده است. میانگین قطعه‌ها در چکیده و عنوان مقاله‌ها ۱۳ و ۲۱۴ است در که جدول آمده است همچنین مقدار واریانس[2] در پرانتز آورده شده است.

جدول ۴–۴ اطلاعات آماری داده‌ها در سطح قطعه

Attribute	Title	abstract
Total	~942k	~15M
Min/max	3/25	85/435
Mean (std)	13.3 (4.8)	214 (65)
Jindex	8.2% (3.6%)	
Overlap	76.63% (17%)	
Total size	~16M title-abstract	

همچنین شباهت‌های آماری برای داده‌ها بین چکیده و عنوان مقاله‌ها محاسبه شد. در اینجا ما از معیار جیندکس[3] و هم‌پوشانی[4] استفاده کردیم. این دو معیار در جدول ۴–۴ آورده شده است. برای محاسبه جینکدس از رابطه ۴–۱ استفاده شده است. متغیر T و A به ترتیب مجموعه قطعه‌های عنوان و چکیده هستند. لازم به ذکر است برای محاسبه معیارهای شباهت همه علائم نگارشی مانند ویرگول و نقطه حذف

[1] Token
[2] Variance
[3] Jindex
[4] Overlap

شده‌اند. هر چه مقدار خروجی این معیار بیشتر باشد نشان‌دهنده‌ی این است که عنوان و چکیده قطعه‌های مشترک بیشتری دارند. همانطور که در جدول ۴–۴ آورده شده است؛ میانگین جیندکس برای تمام داده‌ها مقدار ۸.۲ درصد است. از آنجایی که تعداد کلمات متن چکیده تقریبا بیش از ده برابر تعداد کلمات عنوان است این مقدار منطقی است که نزدیک عدد ۱۰ باشد. از آنجایی که معیار بدست آمده نزدیک به ده است پس شباهت بالایی وجود دارد.

رابطه ۴–۱ جیندکس به عنوان معیار شباهت

$$j(A,T) = \frac{|T \cap A|}{|T \cup A|} = \frac{|T \cap A|}{|T| + |A| - |T \cap A|}$$

همچنین هم‌پوشانی با رابطه ۴–۲ محاسبه شده است. دراین‌رابطه متغیرهای T و A به‌عنوان و چکیده‌های مقاله‌ها را نشان می‌دهد. در این معیار، هم‌پوشانی قطعه‌های خرجی با ورودی بررسی می‌شود. عدد بزرگ‌تر در این معیار نشان‌دهنده شباهت بیشتر است. همان‌طور که در جدول ۴–۴ مشاهده می‌شود میانگین این معیار برای کل جفت چکیده و عنوان‌ها مقدار ۷۶.۶۳ درصد است که شباهت قابل قبول جفت چکیده و عنوان را می‌رساند.

رابطه ۴–۲ همپوشانی به عنوان معیار شباهت

$$O(A,T) = \frac{|T \cap A|}{|T|}$$

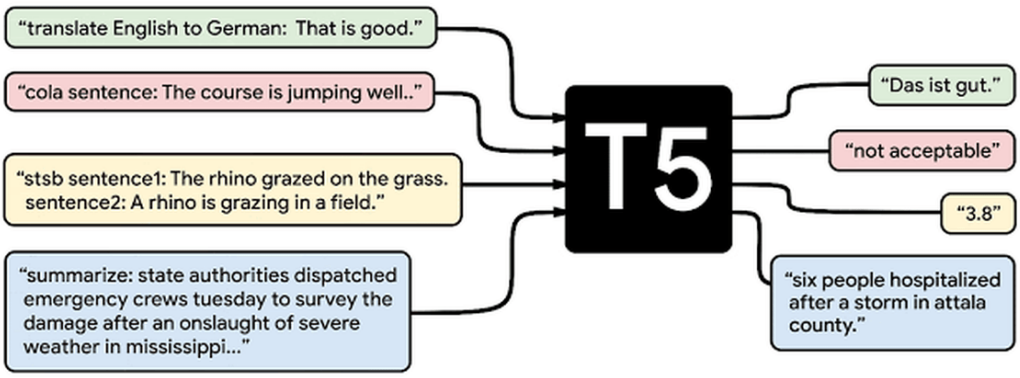
۴-۴ مدل انتقالی mT5

مدل زبانی mT5[1] [48] ورژن چندزبانه‌ی مدل انتقالی T5 [44] و تنها تفاوت این مدل با T5 استفاده پیکره‌گان چند زبانه برای آموزش این مدل است. این مدل ساختار رمزنگار-رمزگشا[2] دارد و می‌تواند کارهای زیر را انجام دهند.

[1] Multilingual Text-to-Text Transfer Transformer or Multilingual T5
[2] Encoder-Decoder

- **مدل زبانی** پیش‌بینی کلمه بعدی در ادامهٔ دنباله‌ای از کلمات
- **بازچینی** کلمه‌های به‌هم‌ریخته را می‌تواند به‌صورت معنادار در جای خود بنشاند
- **پیش‌بینی جای خالی** جای خالی در دنباله‌ای از کلمه‌ها را با کلمه‌ای پر می‌کند

همچنین این مدل با استفاده از مهارت‌های اولیه‌ای که نام‌برده شد می‌تواند اعمال پایین‌دستی[1] پردازش زبان‌های طبیعی مانند خلاصه‌سازی و ترجمه را به طریق متن به متن[2] انجام دهد [54]. شکل ۴–۳ نمایی از ساختار متن‌به‌متن را نمایش می‌دهد. در این طریقه ورودی و خروجی به صورت مستقیم به فرمت متن است. برای مشخص کردن عمل مورد نظر اسم عمل پایین‌دستی را به اول متن ورودی اضافه میکنیم. به عنوان مثال برای دریافت خلاصه از خروجی می‌توان به اول متن «summerize:» اضافه کرد.



شکل ۴–۳ ساختار یکپارچه مدل T5 برای اعمال پایین دستی مرجع [44]. در این شکل فرمت ساختار متن به متن این مدل برای اعمال پایین‌دستی مانند: قابل قبول بود ساختار زبانی (قرمز)، خلاصه‌سازی (آبی)، ترجمه انگلیسی به آلمان (سبز) و تشابه جمله‌ای (زرد)، را نشان می‌دهد .

در این کار، قلب عنوان‌ساز پیاده‌سازی شده مدل mT5-base است که وظیفه‌ی تولید عنوان از متن ورودی را دارد. همانطور که اشاره شده، برای بکارگیری مدل نیاز به تنظیم این مدل است، پس مجموعه دادگان را برای این کار آماده کردیم تا با آن مدل را با همان وزن‌های از پیش‌آموزش داده‌شده‌ی پیش‌فرض،

[1] Down-Stream
[2] Text-to-Text

آموزش دهیم و برای کار خود بهینه کنیم.

برای مرحله آموزش و ارزیابی نیاز است که داده‌ها آزمون و اعتبارسنجی را از جدا و معین کنیم. بعد از به‌هم‌زدن ترتیب[1] داده‌ها ۰.۰۳ درصد از داده‌ها به اعتبارسنجی اختصاص داده شد. دسته دادگان اعتبارسنجی برای تنظیم پارامترها برای مرحله آموزش و محک زدن دقت نهایی مدل استفاده می‌کنیم. در اینجا ورودی چکیده‌ی مقالات و خروجی عنوان متناظر آن مقاله است. قرار است با روش بانظارت جفت نمونه‌ها تعیین شده مدل آموزش ببیند. همچنین ۰.۰۳ درصد داده‌ها به عنوان داده‌های نادیده از مدل برای ارزیابی نهایی مدل‌ها استفاده شد.

برای مرحله آموزش، پس از آزمون و خطاهای متعدد روی داده‌های اعتبارسنجی، سعی شد بهترین فرا پارامترهای برای این پیاده‌سازی انتخاب شوند. مقادیر فراپارامترها در جدول ۴–۶ جدول مقدار فراپارمترهای در نظر گرفته شده برای آموزش مدل آورده شده است. لازم به ذکر است، برای تابع هزینه، مدل mT5 در مرحله آموزش برای همه اعمال پایین‌دستی از یک تابع هزینه استفاده می‌کند، و ما این تابع را تغییر ندادیم و همان وضعیت پیش‌فرض را حفظ کردیم.

جدول ۴–۵ جدول مقدار فراپارمترهای در نظر گرفته شده برای آموزش مدل

فراپارامتر	مقدار	توضیح
نرخ یادگیری	$۱۰^{-۴}$	باتوجه‌به تعداد بسیار زیاد پارامترها برای جلوگیری از بیش‌برازش نیاز است ضریب به‌اندازه کافی کوچک در نظر گرفته شود.
ضریب نرخ یادگیری	۰.۰۱	بعد از هر بار آموزش نرخ یادگیری کمتر می‌شود تا سرعت همگرا شدن به هدف بیشتر شود
دست‌گرمی	۱۰۰	به‌تدریج مقدار نرخ یادگیری از نزدیک صفر به تعداد تعداد گام تنظیم شده صعود می‌کند و بعد از آن برنامه نرخ یادکیری تنظیم شده اعمال می‌شود.
تعداد دور[2]	۳	با تعداد دورهای مختلف به دقت‌ها و خروجی‌های مختلفی می‌توان دست پیدا کرد.

[1] Shuffeling
[2] Epoch

تعداد دسته[1] آموزشی	۲۰ نمونه	بهتر است بعد از هر ورودی آپدیت وزن‌ها انجام شود. دقت خوبی حاصل می‌شود ولی مدت بیشتری برای آموزش صرف می‌شود.
دوره ارزشیابی	هر ۴۰۰ گام آموزشی	بعدازاین تعدادی گام آموزشی می‌توان دقت مدل را ارزیابی کرد.
دسته‌های پردازش موازی	۲	به‌صورت موازی بر روی هسته‌های کارت گرافیک آموزش داده می‌شود. باعث افزایش سرعت آموزش می‌شود.
تعداد حداکثر کلمات ورودی	۵۱۲ قطعه	با توجه به تعداد کلماتی که می‌تواند چکیده‌ها داشته باشند.
تعداد حداکثر کلمات خروجی	۶۴ قطعه	با توجه به تعداد کلماتی که می‌تواند عنوان‌ها داشته باشند.

همچنین برای آموزش از یک کارت گرافیک Tesla p100-16gig کمک گرفته شد و فرایند آموزش تقریباً چهار ساعت برای هر دور طول می‌کشد.

۱-۴-۴ نحوه‌ی اعمال ورودی و دریافت خروجی در مدل mT5

همانطور که گفته شد مدل T5 به صورت متن به متن ارائه شده است (شکل ۳–۵) و در مرحله‌ی آموزش و استفاده از مدل تنها نیاز است حداکثر طول ورودی و خروجی تعیین شود. به عنوان مثال در این کار طول ورودی و خروجی با توجه به طول چکیده‌ها و عنوان‌ها به ترتیب ۵۱۲ و ۶۴ کلمه در نظر گرفته شده است. در مرحله‌ی ورودی در صورتی که تعداد قطعه‌های ورودی کمتر از حداکثر تعیین شده باشد در انتهای کلمات ورودی باید توکن مخصوص <pad> اضافه کرد تا ورودی به حداکثر تعیین شده برسد. در ادامه مدل برای هر یک از قطعه‌ها یک جاسازی از پیش تعیین‌شده در نظر می‌گیرد. در این راستا بخش رمزگشای مدل mT5 وظیفه دارد در هر مرحله با استفاده از مکانیزم توجه، جاسازی‌های قطعه‌های ورودی را بهتر و بهتر کند تا این جاسازی را به رمزگشا ارائه دهد. این مدل در خروجی کلمات را به صورت one-hot رمز

[1] Batch

گشایی می‌کند؛ پس در لایه‌ی آخر مدل به اندازه تعداد کلمات مجموعه واژگان مدل نورون وجود دارد و هر یک نماینده یک کلمه است. برای تولید اولین کلمه در خروج، رمزنگار توکن <start> را به خود ورودی می‌دهد و با استفاده از جاسازی‌های تولید شده در مرحله‌ی رمزگشا، یک کلمه را در لایه آخر پیشنهاد می‌دهد. در ادامه دوباره رمزگشا کلمه‌ها پیشین گذشته تولید شده را به خود ورودی می‌دهد تا کلمه‌ی جدید در خروجی پیشنهاد داده شود. این کار تا زمانی انجام می‌شود که در خورجی قطعه‌ی <end> پیشنهاد داده شود و این به این معنی است که رمزگشا تمام کلمات را در کنار هم گذاشته تا خروجی مناسب تولید شود.

پیش‌پردازش و پس‌پردازش‌ها برای مدل mT5

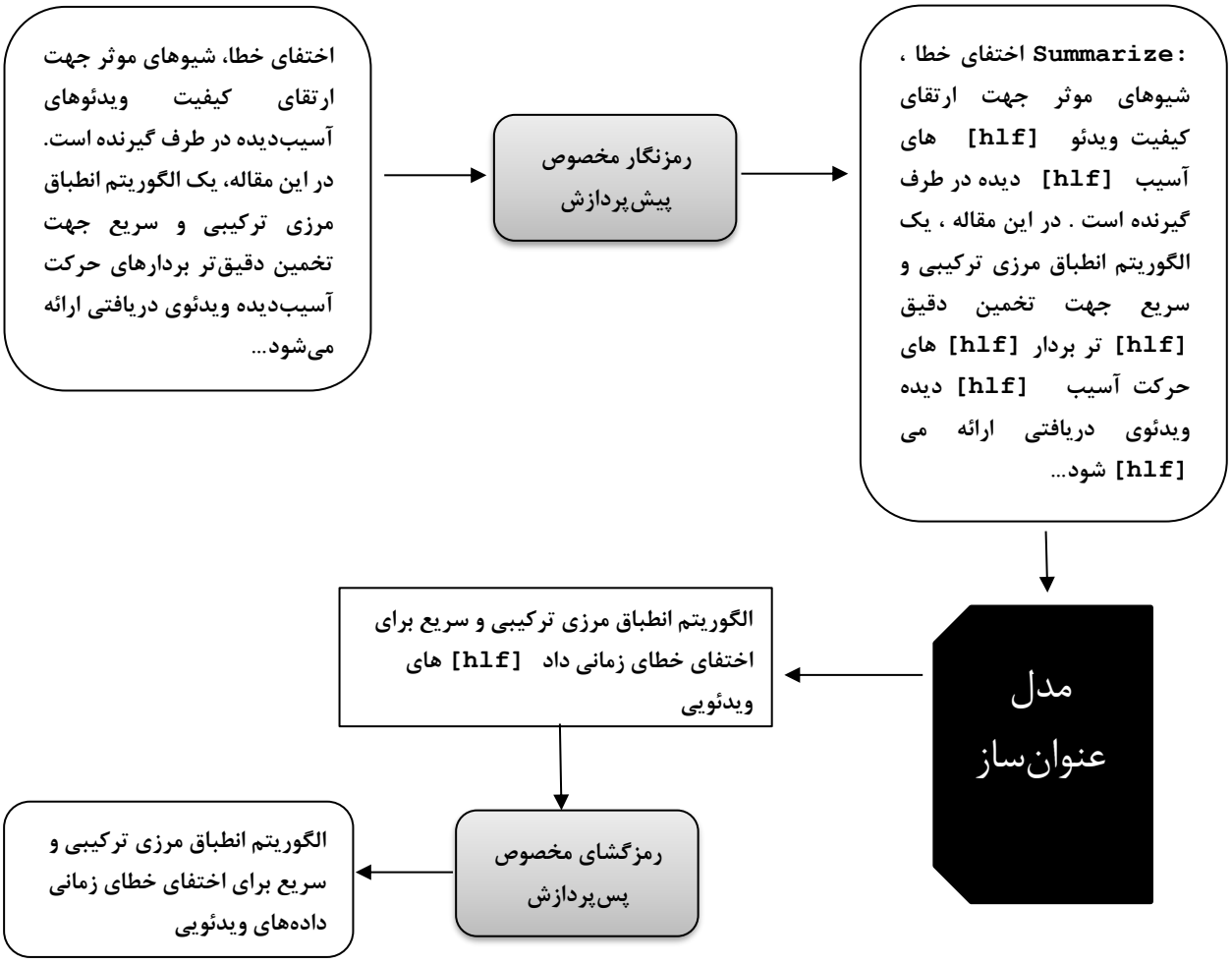
دو پیش‌پردازش قبل از ورودی دادن متن موردنظر به مدل برای عمل پایین‌دستی خلاصه‌ساز (به‌صورت دقیق‌تر در کار ما عنوان‌ساز) هم برای انجام آموزش مدل و هم برای استفاده از مدل آموزش‌داده‌شده در نظر گرفته شده است.

پیش‌پردازش اول اضافه‌کردن رشته [1] "summarize: " در ابتدای متن ورودی به مدل داده می‌شود. در واقع با این کار، همان‌طور که در بخش شبکه انتقالی T5 گفته شد، یک نشانه[2] برای ساختار متن به متن مدل است تا عمل خلاصه‌سازی بر روی ورودی انجام دهد. البته لازم به ذکر بدون آموزش عمل ورودی، خروجی متن نامفهوم است و ارزش ندارد. در بخش بعدی مقایسه در زبان فارسی و انگلیسی بدون آموزش انجام خواهد شد.

همچنین نیم‌فاصله‌ها برای زبان فارسی مسئله‌ای است که باید آن را در نظر گرفت. به‌عنوان‌مثال «فناوری‌ها» یک قطعه به‌حساب می‌آید ولی ما برای بهبود نتایج مدل، این کلمه را سه قطعه «فناوری»،

[1] String
[2] Flag

«hlf» و «ها» در نظر می‌گیریم. در اینجا منظور از «hlf» نیم‌فاصله است. با این کار ما ساختار جمع (و دیگر جاهایی که نیاز به نیم‌فاصله است) را به مدل آموزش می‌دهیم و مدل توانایی خروجی دادن نیم‌فاصله در جاهای مناسب را بدست می‌آورد. طبیعی است در خروجی نهایی مدل کاراکتر «hlf» را توسط یک رمزگشا ساده به نیم‌فاصله تبدیل می‌کنیم. در شکل ۴–۴ فرآیند پیش‌پردازش برای مدل انتقالی mT5 توضیح داده شده است.



شکل ۴–۴ دیاگرام پیش‌پردازش و پس‌پردازش‌های مورد نیاز برای مدل mT5. اضافه کردن قطعه "summarize:" در اول متن ورودی و همچنین تبدیل کاراکتر U+200C (نیم فاصله) به کاراکتر «hlf» و جدا کردن آن از قطعه‌های طرفین. همچنین در خروجی یک دیکودر قرار داده شده تا کاراکتر «hlf» را به U+200C برگرداند و با کلمه‌های اطراف یکپارچه کند تا به یک قطعه تبدیل شود.

۴-۵ ماژول نویززدا

بعد از تست و بررسی عنوان‌های تولید شده توسط مدل mT5 که در بخش قبلی بررسی شد، بعضی مواقع در عنوان‌ها کلمات نامفهوم مشاهده می‌شد و یا عنوان با کلمه‌ای نامناسب پایان می‌یافت. ما در اینجا این نوع کلمات را نویز تلقی می‌کنیم و تصمیم گرفتیم که برای رفع این مشکل یک ماژول اضافی در دنباله مدل mT5 قرار دهیم تا عنوان‌هایی بدون ایراد را خروجی بگیریم. برای این کار یک مدل mT5-small به کار گرفته شد و آن را برای نویززدایی دوباره آموزش دادیم و یا تنظیم کردیم. روش آموزش این مدل این است که مدل با دیتاست یاد گرفت تا عنوان تولیدشده توسط ماشین را به عنوان مرجع تولیدشده توسط انسان نگاشت دهد.

در اینجا ورودی مدل ما، عنوان‌های تولیدشده توسط مدل زبانی و خروجی هدف، عناوین مرجع متناظر است. لازم به ذکر است در ابتدای نمونه‌های ورودی رشته‌ی " :denoising" چسبانده می‌شود.

پس از آزمون و خطاهای متعدد با داده‌های اعتبارسنجی، سعی شد بهترین فراپارامترها برای این پیاده‌سازی انتخاب شود. مقادیر فراپارامترها در جدول ۴–۶ جدول مقدار فراپارمترهای در نظر گرفته شده برای آموزش مدل آورده شده است.

جدول ۴–۶ جدول مقدار فراپارمترهای در نظر گرفته شده برای آموزش مدل

فراپارامتر	مقدار
نرخ یادگیری	$۱۰^{-۴}$
ضریب نرخ یادگیری	۰.۰۱
دست‌گرمی	۱۰۰
تعداد دور	۴
تعداد دسته آموزشی	۱ نمونه
دوره ارزشیابی	هر ۲۰۰ نمونه
دسته‌های پردازش موازی	۲

۴-۶ معماری پیشنهادی شماره ۱

شکل زیر یکی از معماری‌های پیشنهاد شده برای پیاده‌سازی خلاصه‌ساز فارسی در این پایان‌نامه می‌باشد که شامل ماژول نویزدایی است که در بخش قبل توضیح داده شد. در این معماری، مدل mT5 به همراه یک نویززدا وظیفه تولید عنوان مناسب را دارند. این دو مستقلا آموزش داده می‌شوند و درنهایت، نویززدا به دنبال مدل انتقالی mT5 قرار داده می‌شود. همچنین در خروجی و ورودی ماژول‌های پیش‌پردازش و پس‌پردازش قرار داده شده است. در ادامه به توضیح هر یک از این ماژول‌های خواهیم پرداخت.

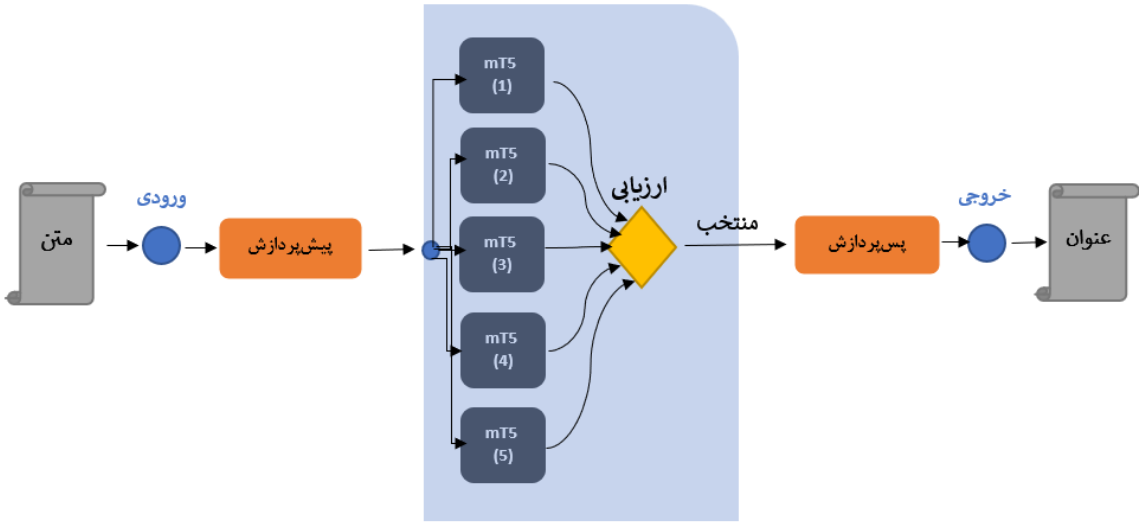
- ماژول پس‌پردازش و پیش‌پردازش بدون تغییر همان فرایندی است که در شکل ۴–۴ نشان داده شده است.
- مدل mT5 با فراپارامترهای جدول ۴–۵ با دادگان پیشنهادی این کار آموزش داده شد.
- مدل نویززدا یک مدل mT5-small است که روند آموزش آن در بخش ۴-۵ ماژول نویززدا توضیح داده شد.

شکل ۴–۵ معماری پیشنهاد شده شماره یک برای عنوان‌ساز فارسی چکیده‌ای، همون طور که مشاهده می‌شود از چند ماژول پیش‌پردازش و پس‌پردازش، مدل انتقالی و نویززدا تشکیل شده است.

۴-۷ معماری پیشنهادی شماره ۲

همانطور که گفته شده خروجی‌های مدل mT5-base تنها که برای عنوان‌سازی آموزش داده‌شده‌بود مستعد کلماتی نامعلوم در خروجی و عنوان‌هایی بود که با کلمات نامربوط تمام می‌شدند. برای رفت این مشکل یک معماری دیگر پیشنهاد داده‌ها شد تا خروجی‌هایی بی نقص تولید شود.

مدل پیشنهادی دوم مشابه معماری گروه مدل‌ها[1] پیاده‌سازی شده است. برای اینکار از پنج مدل زبانی mT5 استفاده شده است. هر یک با دیتاست پیشنهادیمان آموزش داده شده‌اند و تفاوتشان تنها در تعداد دوره‌هایی هستند که آموزش دیده‌اند. این پنج مدل از یک تا پنج دوره آموزش شده‌اند در نتیجه می‌توانیم ۵ خروجی متفاوت از هم داشته باشیم. در این معماری، ورودی به هر یک از پنج مدل داده می‌شود و تا پنج عنوان تولید شود. این عنوان‌ها را کاندید در نظر می‌گیریم تا در مرحله بعد، با معیار شباهت همپوشانی با متن ورودی مقایسه شوند. در نهایت آن عنوانی که شباهت بیشتری با متن مرجع داشته باشد انتخاب و به عنوان خروجی برگردانده می‌شود. در شکل ۴–۶ معماری این مدل آورده شده است. لازم به ذکر است که مرحله‌ی پیش‌پردازش و پس‌پردازش همان مراحل در نظر گرفته شده برای مدل mT5 (شکل ۴–۴) هستند.



شکل ۴–۶ معماری پیشنهادی شماره ۲ برای عنوان ساز چکیده‌ای فارسی

۴-۸ جمع‌بندی

در این فصل برای پیاده‌سازی عنوان‌ساز فارسی استفاده از مدل‌های پیش‌آموزش داده شدهٔ انتقالی استفاده شده و دو معماری برای پیاده‌سازی عنوان‌ساز فارسی چکیده‌ای ارائه شده. برای این کار داده‌های مناسب با

[1] Ensemble models

تنوع موضوعی بالا انتخاب شده و مورد پیش‌پردازش‌های اولیه قرار گرفت. در یک رویکرد مدل mT5 با داده‌ها آماده شده آموزش داده شده و بهمراه یک نویززدا عنوان‌ساز چکیده‌ای ارائه شده است. در رویکرد دیگر، پیشنهاد شود که گروه مدل‌های متفاوت خروجی های مختلفی را به عنوان کاندید برای عنوان نهایی، تولید کنند؛ در نهایت با معیار شباهت عنوان برتر انتخاب و ارائه شود.

فصل ۵ نتایج و آزمایش‌ها

۵-۱ مقدمه

برای پیاده‌سازی معماری‌های پیشنهاد داده شده، ابتدا مدل‌های مورد استفاده جداگانه قبل از اینکه کنار هم چیده‌شوند آموزش داده شدند. در این بخش از روند تنظیم پارامترها و آموزش مدل بحث خواهیم کرد. سپس به بررسی و مقایسه‌ی نتایج خواهیم پرداخت.

۵-۲ خلاصه‌ای از روندتنظیم پارامترها مدل‌ها و آموزش

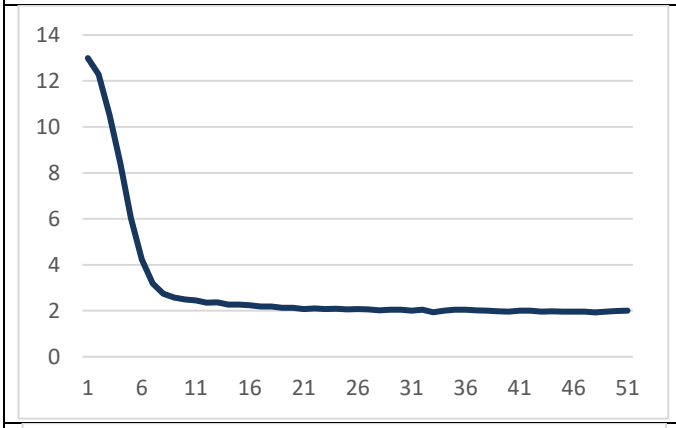
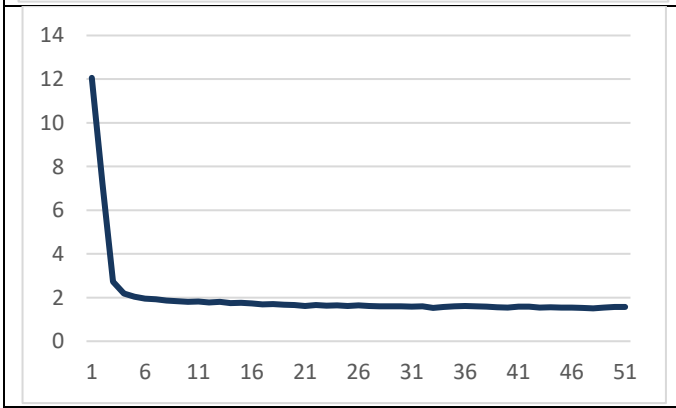
برای تنظیم فراپارامترهای[1] مدل، از آنجایی که مدل mT5 حجیم است (نزدیک ۲.۵ گیگ با ۲۲۰ میلیون پارامتر [44]) حتی زمان اعتبارسنجی با دادگانی متشکل از تنها ۰.۰۳ درصد از کل دیتاست (حدود ۲ هزار نمونه)، تقریبا ۲۰ دقیقه طول می‌کشید. از آنجاییکه منابع محاسباتی برای تنها ۲۴ ساعت آموزش مداوم در اختیار بود؛ به همین خاطر برای تنظیم پارامترها روش‌های جست‌وجو[2] عملا منطقی به نظر نمی‌آمد؛ در نتیجه با آزمون و خطا روی داده‌های اعتبارسنجی سعی شد بهترین پارامترها برای مدل انتخاب شوند. لازم به ذکر است آموزش مدل با استفاده از کتابخانه‌ی پایتورچ[3] در پایتون انجام شده. همچنین بخش‌های دیگر با کتابخانه‌های رایج یادگیری ماشین مانند نامپای[4] با زبان برنامه نویسی پایتون پیاده‌سازی شد.

برای تنظیم پارامتر نرخ‌یادگیری، با در نظر گرفتن اینکه مدل حجیم است، برای جلوگیری از بیش‌برازش[5] نرخ‌یادگیری را $۱۰^{-۴}$ در نظر گرفتیم. در جدول ۵–۱ زیر می‌توان تاثیر نرخ یادگیری بر مقدار ROUGE-1 داده‌های تست و همچنین نمودار خطای آموزش را مشاهده کرد. لازم به ذکر است که نرخ‌یادگیری با رویکرد تطبیقی در هر گام آموزش به صورت خطی بیشتر می‌شود تا سرعت آموزش مدل در جهت نقطه بهینه بیشتر شود. نرخ افزایش خطی نرخ‌یادگیری با هر بار آموزش ۰.۰۱ است. همچنین مقدار

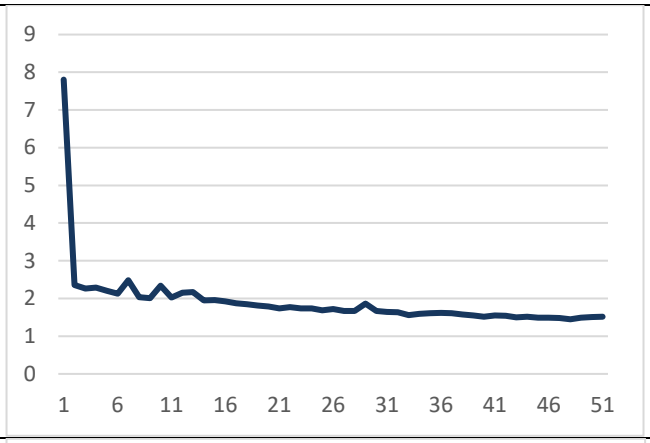
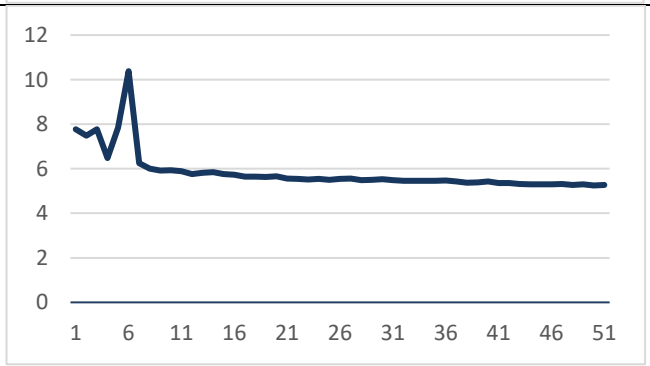
[1] Hyperparameter
[2] Grid search
[3] Pytorch
[4] Numpy
[5] Overfitting

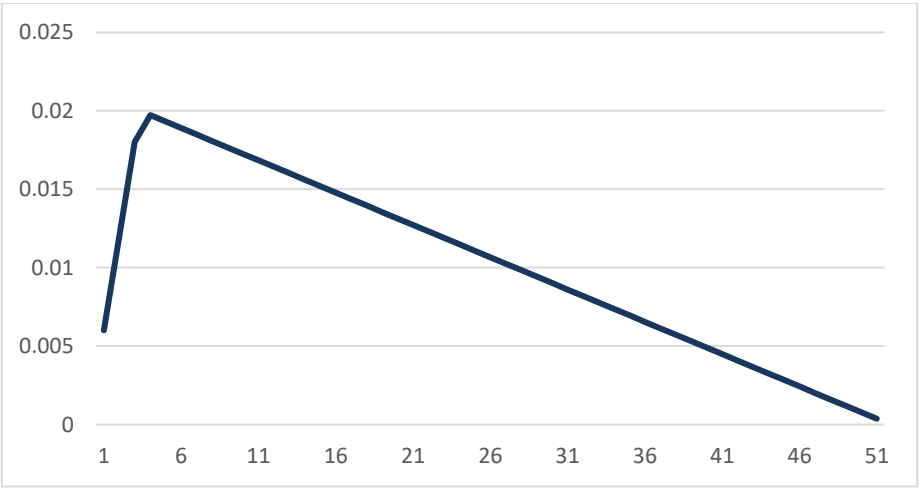
گام‌های‌دست‌گرمی[1] عدد ۱۰۰ انتخاب شده است. با روند دست‌گرمی، نرخ یادگیری در بازه‌ی ۱۰۰ گام آموزشی از مقدار نزدیک صفر به مقدار تعیین شده (در اینجا $۱۰^{-۴}$) می‌رسد و برنامه‌ی روند صعودی نرخ‌یادگیری، همانطور که تنظیم شده است شروع می‌شود. این پارامتر کمک می‌کند تا مدل در ابتدا با تزریق داده‌ها منحرف نشود و برای احتیاط اولین گام‌های آموزش بسیار کوچک باشد. درشکل ۵–۱ روند تغییرات نرخ یادگیری در یک دوره نمایش داده شده است؛ در ستون نمودار خطای آموزش، روند همگرایی مدل با نرخ‌یادگیری مدل نمایش داده شده است این نمودارها خطای آموزش نسبت به هرگام آموزش را نشان می‌دهند. همچنین تعداد گام‌های آموزش ۵۱ بار است. این مقدار از تقسیم تعداد داده‌ها و دسته‌ها ضربدر تعداد پردازش‌های موازی، به دست می‌آید.

جدول ۵–۱ تاثیر نرخ یادگیری بر معیار ROUGE1

نرخ‌یادگیری	ROUGE-1	نمودار خطای آموزش loss/step
$۱۰^{-۵}$	۴۱.۱۷	
$۱۰^{-۴}$	**۴۳.۳۸**	

[1] Warmup steps

	۴۲.۷۴	$۱۰^{-۳}$
	۰	$۱۰^{-۲}$



شکل ۵–۱ روند تغییرات نرخ‌یادگیری در یک دوره

برای تنظیم اندازه دسته[1] طبق جدول ۵–۲ مقدار مناسب انتخاب شد. مشاهده می‌شود آموزش مدل با بدون دسته‌بندی دقتی بهتر از سایر ارائه می‌دهد ولی زمان آموزش آن می‌تواند دو تا سه برابر بیشتر از

[1] Batch

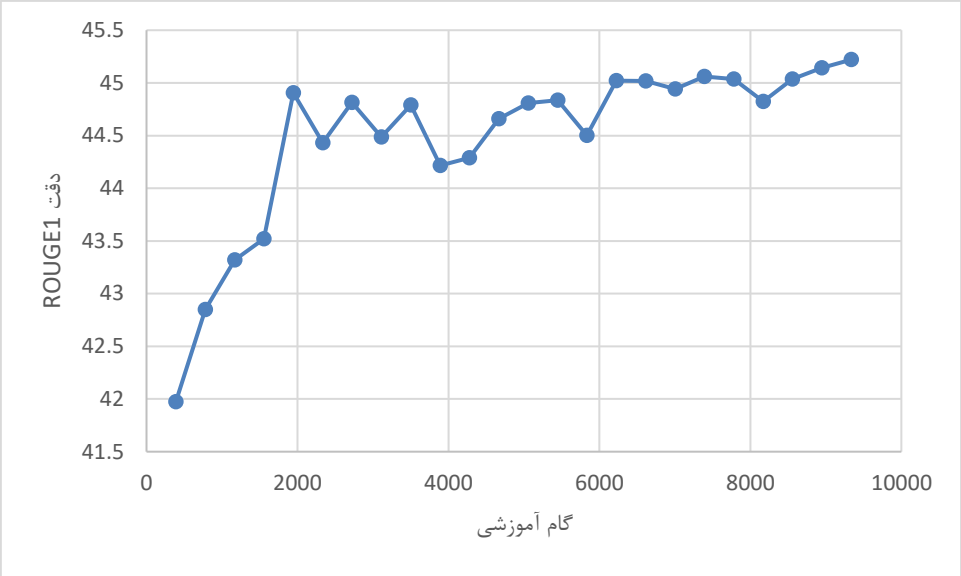
آموزش با دسته‌هایی ۲۰ تایی یا ۱۰ تایی باشد. به همین دلیل تصمیم گرفته شد برای آموزش از دسته های آموزش ۲۰تایی استفاده کنیم تا علاوه بر داشتن دقت قابل قبول زمان آموزش هم برای آزمون و خطاهای بیشتر مناسب باشد. لازم به ذکر است که برای صرفه‌جوی در زمان مدل به صورت موازی آموزش داده شد. اینطور که در هر بار بروزرسانی، دو نمونه جداگانه به مدل همزمان ورودی داده می‌شود و تغییرات وزن‌ها در بروزرسانی در مرحله انتشاربه‌عقب[1]، از میانگین گرادیان این دو مدل محاسبه می‌شود.

جدول ۵–۲ دقت مدل در دسته‌های آموزش مختلف (نرخ یادگیری ۰.۰۰۰۱)

تعداد نمونه‌های هر دسته آموزشی	دقت در یک دور(ROUGE1)	مدت زمات یک دور آموزش (ساعت)
۱	۴۳.۹۲	~۶
۱۰	۴۳.۱۷	~۴.۵
۲۰	**۴۳.۳۸**	**~۲.۵**

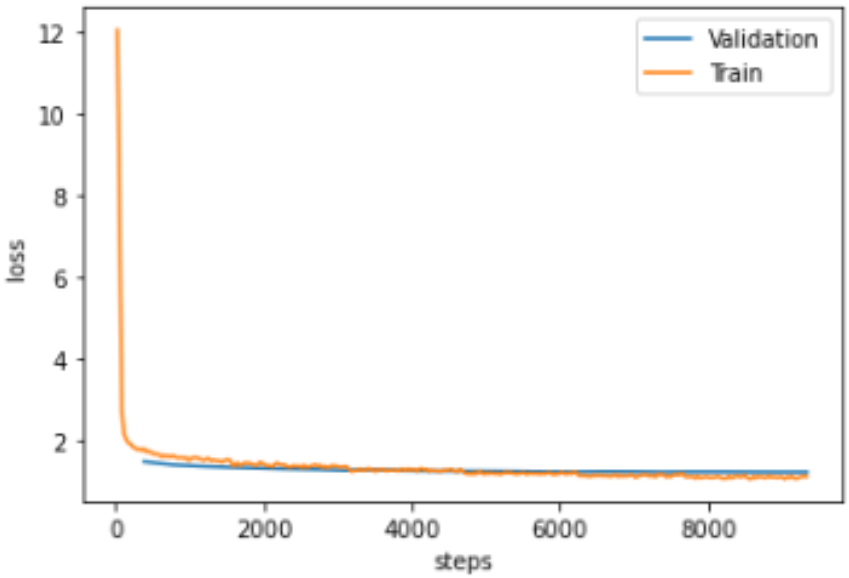
برای اندازه‌گیری تعداد دوره‌های مناسب از ۱ تا ۶ دوره مدل mT5-base را با داده‌های آموزشی خود تنظیم کردیم. آموزش با تنها ۱ دور، نتایج قابل قبولی (دقت ۴۳.۳۸) برای داده‌های اعتبارسنجی بدست آمد و همچنین خروجی‌های خوبی مشاهده شد. در شکل ۵–۲ دقت مدل برای ۶ دوره و یا ۹۳۳۶ گام آموزش (دسته‌هایی شامل ۴۰ نمونه) قابل مشاهده است. در ادامه برای مقایسه معماری‌ها پیشنهاد شده با نتایج این مدل از مدلی استفاده شد که در آخرین دوره یعنی دوره‌ی ۶ (دقت ۴۵.۲۲) بدست آمده است.

[1] Backpropogation



شکل ۵–۲ نمودار دقت بر تعداد دوره‌های آموزشی

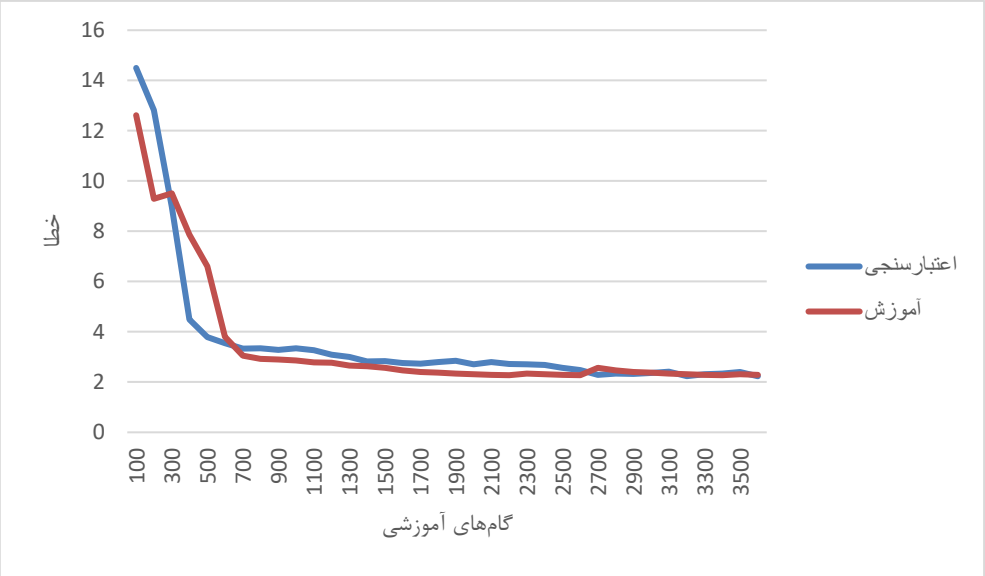
در شکل ۵–۳ نمودار مقدارتابع خطا بر حسب هر گام‌های (هر گام با دسته ۴۰ تایی نمونه‌ها آموزش داده می‌شود.) آموزشی در ۶ دوره آورده شده است. همانطور که مشاهده می‌شود در ۶ دوره آموزش، همچنان بیش‌برازش مشاهده نمی‌شود و حتی با افزایش دوره‌ها دقت ROUGE1 مدل با داده‌های اعتبارسنجی افزایش هم پیدا می‌کند (شکل ۵–۲). از آنجایی که خروجی مدل‌ها در هر دوره ممکن است فرق داشته باشد، ما از این خاصیت برای استفاده در معماری پیشنهاد شده‌ی شماره ۲ (5*mT5+selector) بهره بردیم. همچنین برای معماری شماره ۱ برای مدل اصلی که عنوان تولید می‌کند از mT5-base استفاده شد که با ۶ دوره تنظیم شده است.



شکل ۵–۳ نمودار خطار آموزش برای گام‌های آموزشی به همراه خطای داده‌های اعتبارسنجی

همچنین برای مدل پیشنهادی شماره ۱ (mT5+denoiser) شکل ۴-۵، از یک ماژول نویززدا استفاده شده است. این ماژول یک مدل mT5-small است که با روند زیر برای عمل نویززدایی عنوان‌های تولید شده در بخش قبل، تنظیم شده است.

- مدل mT5-small برای کار عنوان‌ساز با داده‌های پیشنهادی، آموزش داده شد.
- برای آموزش مدل برای کار نویززدایی، مدلی که از مرحله قبل آموزش داریم را با دادگانی مخصوصی که برای اینکار تهیه کردیم آموزش دادیم. دادگان متشکل از داده‌های آزمونی است که برای آموزش مدل در مرحله قبل جدا شده بود. لذا، این داده‌ها توسط مدل دیده نشده است. برای این کار، داده‌های آزمون را به مدل ورودی دادیم تا عنوان‌ها تولید شوند. سپس عنوان‌های تولید شد و عنوان‌ها مرجع را جفت کرده و مدل را با این مجموعه داده‌ی جدید برای کار نویززدایی آموزش می‌دهیم. لازم به ذکر است پارامترهای مورد استفاده در اینکار در جدول ۴–۶ آورده شده است. در شکل ۵–۴ روند مقادیر خطا در گام‌های آموزش برای داده‌های اعتبارسنجی آورده شده است.



شکل ۵–۴ نمودار خطا در هر گام آموزشی مدل نویززدا

۵-۳ نتایج و مقایسه‌ها

در این بخش معماری‌های پیشنهاد شده مورد ارزیابی قرار می‌گیرند و نتایج خروجی آن‌ها را بررسی می‌کنیم. همچنین در ادامه به مقایسه هر یک می‌پردازیم.

۵-۳-۱ معماری پیشنهادی شماره ۱ (mT5+denoiser)

این معماری از دو مدل زبانی mT5 تشکیل شده است. همانطور که درشکل ۵–۵ مشاهده می‌شود، متن ورودی پس از پیش‌پردازش به مدل mT5 برای عمل خلاصه‌سازی ورودی داده می‌شود تا یک عنوان تولید شود؛ سپس به دنبال آن برای بهبود عنوان تولید شده از مدل نویززدا استفاده می‌شود تا در خروجی عنوانی بهتر ارائه شود. مشخصا این مدل‌ها قبل از کنار هم چیده شدن از قبل برای کارهای موردنظر آموزش دیده شده اند. در جدول ۵–۳ دو نمونه از عنوان‌های تولید شده توسط پیشنهادی شماره ۱ نمونه‌هایی از عنوان تولیدهای شده توسط این مدل ارائه شده است. همچنین عنوان تولیدی توسط مدل mT5 تنها (با ۶ دوره

آموزش)، بدون نویززدا هم آورده شده است. لازم به ذکر است، متون ورودی از داده‌های آزمون انتخاب شده‌اند. دقت این مدل پیشنهادی برای عنوان‌سازی در جدول ۵–۴ آورده شده است.

شکل ۵–۵ معماری پیشنهادی شماره ۱ برای عنوان‌ساز فارسی

جدول ۵–۳ دو نمونه از عنوان‌های تولید شده توسط پیشنهادی شماره ۱

#	نمونه‌ها
۱	**متن چکیده (abstract):** هدف پژوهش حاضر مقایسه فرایندهای شناختی (حافظه کاری و بازداری پاسخ) در افراد وابسته به متآمفتامین بود. پژوهش حاضر از نوع علی–مقایسه‌ای است. جامعه پژوهش کلیه افراد وابسته به متآمفتامین و افراد غیروابسته به مواد در شهر ایلام بودند. روش نمونه‌گیری پژوهش حاضر به شیوه دردسترس و نمونه پژوهش شامل ۳۰ نفر وابسته به متآمفتامین، و ۳۰ نفر از افراد عادی بود. از آزمون رنگ-واژه استروپ (نوع رایانه‌ای) برای ارزیابی بازداری پاسخ و برای ارزیابی حافظه کاری نیز خرده آزمون فراخنای ارقام حافظه وکسلر (نوع رایانه‌ای) مورد استفاده قرار گرفت. داده‌ها با روش آماری واریانس چندمتغیره تجزیه و تحلیل شد. نتایج نشان داد عملکرد افراد وابسته به متآمفتامین درمؤلفه‌های فراخنای ارقام رو به جلو و فراخنای ارقام معکوس با گروه به‌هنجار تفاوت معناداری دارد. همچنین عملکرد افراد وابسته به متآمفتامین در بازداری پاسخ تفاوت معناداری با گروه به‌هنجارداشت ولی در مؤلفه‌های زمان واکنش همخوان و زمان واکنش ناهمخوان و مؤلفه‌های بدون پاسخ ناهمخوان تفاوت معناداری بین دو گروه مشاهده نشد. نتایج پژوهش نشان داد که وابستگی به متآمفتامین تأثیر معناداری بر حافظه کاری و بازداری پاسخ افراد وابسته دارد.
	عنوان مرجع (reference title): ارزیابی عملکرد حافظه کاری و بازداری پاسخ در افراد وابسته به متآمفتامین و افراد عادی
	عنوان تولیدشده (mT5): مقایسه فرایند های شناختی در افراد وابسته به مت آمفتامین افراد
	عنوان تولیدشده (mT5+denoise): مقایسه فرایند های شناختی در افراد وابسته به مت آمفتامین
۲	**متن چکیده (abstract):** سمیت تنفسی اسانس گیاهان رازیانه Foeniculum vulgare Miller کلپوره Teucrium polium Boiss و مرزه Satureja hortensis L روی سوسک چهار نقطه ای حبوبات Callosobruchus maculatus F. در یک دوره زمانی ۲۴ ساعته مورد بررسی قرار گرفت اسانس ها با استفاده از دستگاه کلونجر به روش تقطیر با آب

تهیه شد آزمایش ها در شرایط دمای ۲۸±۲ C° و رطوبت نسبی ۶۰±۵ درصد و در تاریکی انجام شد از هر اسانس ۶ غلظت در ۶ تکرار مورد آزمایش قرار گرفت شناسایی ترکیبات اسانس ها با استفاده از روش GC-MS صورت گرفت نتایج نشان داد که اصلی ترین ترکیبات در اسانس رازیانه ترانس انتول ۶۱/۶۰٪ و فنچون ۱۴/۱۲٪ و در اسانس کلپوره پیپریتنون اکساید ۷۷/۲۱٪ آلفا پینن ۳۳/۱۱٪و کارون ۲۹/۱۱ ٪ بود کارواکرول ۱۳/۵۰٪ و تیمول ۷۷٪/۲۷ عمده ترین ترکیبات اسانس مرزه بودند هر سه اسانس روی حشرات کامل سمیت تنفسی بالایی داشتند مرگ و میر حشرات کامل یک روزه در هر دو جنس نر و ماده با افزایش غلظت اسانس ها افزایش یافت.
عنوان مرجع (reference title): بررسی سمیت تنفسی اسانس سه گیاه دارویی روی حشرات کامل سوسک چهار نقطه ای حبوبات Callosobruchus maculatus F. (Coleoptera: Bruchidae)
mT5: بررسی سمیت تنفسی اسانس گیاهان رازیانه کلپوره رو
عنوان تولیدشده (mT5+denoise): بررسی سمیت تنفسی اسانس گیاهان رازیانه و کلپوره

همانطور که در جدول ۵–۳ مشاهده می‌شود. خروجی تولیدشده به خوبی می‌تواند عنوانی را برای متن ورودی ارائه دهد. همچنین قسمت نویززدا توانسته است عنوان‌هایی که دارای قطعه‌های نامربوط در انتها هستند را حذف کند. وجود قطعه‌های نامربوط در عنوان‌های تولیدی توسط مدل mT5-base، یکی از مشکلاتی بود که حتی با افزایش تعداد دورها برای آموزش هم حل نمیشد. در همین راستا تصمیم به طراحی یک نویززدا برای خروجی این مدل گرفتیم.

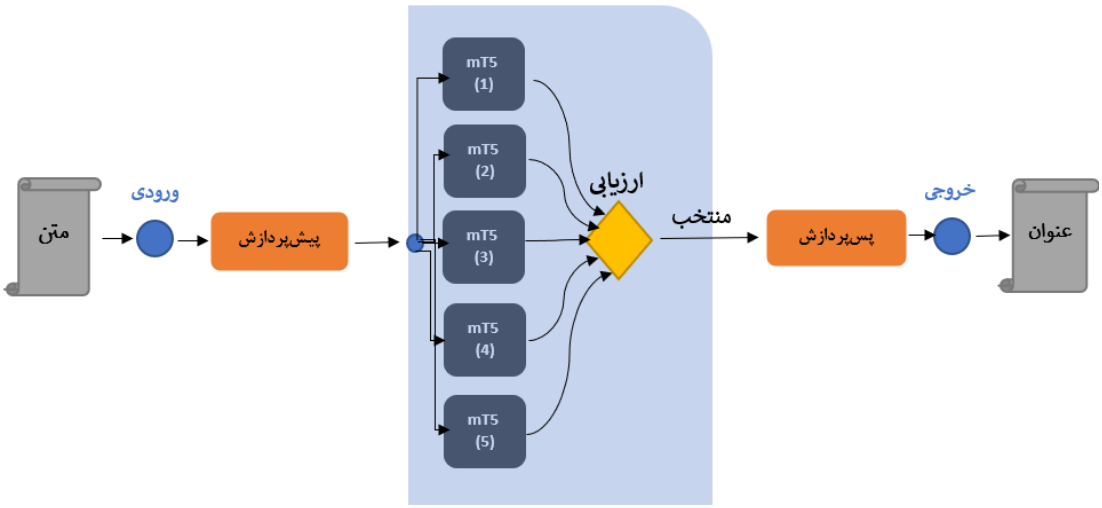
جدول ۵–۴ دقت مدل پیشنهادی۱ به همراه دقت مدل عنوان‌ساز mT5

مدل	ROUGE-L	ROUGE-2	ROUGE-1
mT5	۴۱.۷۲	۲۷.۴۶	۴۵.۰۳
پیشنهادی شماره ۱ (mT5+denoiser)	**۴۲.۶۵**	**۲۸.۲۳**	**۴۵.۹۵**

۵-۳-۲ معماری پیشنهادی شماره ۲ (5*mT5+selector)

همانطور که گفته شد، مدل mT5 به تنهایی برای کار ما، ممکن است خروجی‌هایی همراه با قطعه‌های نامعلوم تولید کند. در معماری پیشنهادی شماره ۱، توسط یک مدل نویززدا سعی شد خروجی‌ها بهبود یابند. در ادامه به توضیح معماری دیگری برای رفع این مشکل خواهیم پرداخت. در معماری شماره ۲، ورودی همزمان به چندین مدل mT5-base داده می‌شود که هرکدام از ۲ تا ۶ اپوک برای عنوان‌سازی

تنظیم شده‌اند. اینطور ۵ عنوان متفاوت داریم که می‌توان بهترین را از بین آن‌ها انتخاب کنیم. برای انتخاب به صورت اتوماتیک، ما از معیار شباهت هم‌پوشانی (رابطه ۴–۱) استفاده می‌کنیم و شباهت عنوان‌های تولیدی را با متن مرجع می‌سنجیم. آن عنوانی که عدد بیشتری را از این معیار کسب کند، برای عنوان نهایی انتخاب می‌شود و خروجی داده می‌شود. در جدول چند نمونه از عنوان‌های تولیدی توسط این مدل خواهیم پرداخت. همچنین در جدول ۵–۶ دقت مدل پیشنهادی در این بخش آورده شده است.



شکل ۵–۶ معماری پیشنهادی شماره ۲ برای عنوان ساز چکیده‌ای فارسی

جدول ۵–۵ دو نمونه از عنوان‌های تولید شده توسط پیشنهادی شماره ۲

#	نمونه‌ها
۱	**متن چکیده (abstract):** دوران صفویه یکی از دوران‌های درخشان ایران است . این دوران از چند جهت اهمیت دارد اول اینکه بعد از قرن ها مجددا یک دولت واحدی بر سرزمین های ایران زمان ساسانی تشکیل شد و به تعبیر دیگر ملیت ایرانی شکل گرفت و دوم اینکه از جهت مذهبی ، شیعه در این سرزمین رسمیت یافت که این رسمیت تاکنون ادامه دارد . اما از جهت دیگر رشد و آبادانی اقتصادی و فرهنگی و اجتماعی و علمی در این دوره را شاهد هستیم . اما همین حکومت در برابر عده ای شورشی افغان سر تسلیم فرود آورد و سقوط نمود . در اواخر دوره صفویه آسیب‌های مختلف اجتماعی و سیاسی در قلمرو صفویه وجود داشت که زمینه‌های شکست و اضمحلال حکومت صفویه را فراهم نمود . برخی از دانشمندان و متفکران در آن روزگار با توجه به تحولات اجتماعی و سیاسی صفویه آثاری را در آسیب‌های اجتماعی آن دوران به رشته تحریر در آوردند و برای حل آن آسیب‌ها راه حل‌هایی ارایه داده بودند . قطب الدین نیریزی(۱۱۰۰-۱۱۷۳) از جمله این افراد است که رساله‌های مختلفی را در نقد اوضاع صفویه نگارش نموده است و راه حل هایی را نیز برای اصلاح جامعه پیشنهاد کرده است . قطب الدین نیریزی علت اصلی آسیب های اجتماعی را

	در ساختار حاکمیت صفویه می داند .و حل مسائل اجتماعی و آسیب های اجتماعی را در اصلاح ساختار حکومت می داند .
	عنوان مرجع (reference title): تأثیر ساختار حکومت بر آسیب های اجتماعی وسیاسی دوره صفویه از نگاه قطب الدین نیریزی
	عنوان تولیدشده (mT5): آسیب های اجتماعی در قلمرو صفویه
	عنوان تولیدشده (معماری شماره ۲): بررسی نقش آسیب‌های اجتماعی و آسیب‌های اجتماعی در قلمرو صفویه
۲	**متن چکیده (abstract):** شاخص های رشدی نوزادان از موارد مهم در ارزیابی و مراقب تهای دوران بارداری است که تأثیر بسزایی در زندگی پس از تولد متولدین دارد. ازاین رو، شناخت عوامل مرتبط با آن امری مهم و ضروری است. ازاین رو پژوهش حاضر با هدف، بررسی ارتباط سبک زندگی مادران باردار بر شاخص های رشد جسمانی نوزادان پس از تولد انجام گرفت. این پژوهش مراجعه کننده به بیمارستان /۱±۵/ همبستگی توصیفی – مقایسه ای بر روی ۴۱۲ زن باردار (با میانگین سنی ۲۸ سال) مهدیهٔ تهران به صورت مقطعی و میدانی انجام گرفت. افراد به روش در دسترس انتخاب شدند. گردآوری داده ها با و اطلاعات مربوط به شاخص های رشدی نوزادان از پروندهٔ پزشکی موجود در (LSQ) استفاده از پرسشنامهٔ سبک زندگی مقایسه شد. آزمون NCHS بیمارستان مهدیهٔ تهران استخراج و ثبت شد. سپس شاخ صهای رشدی با نورم استاندارد مستقل و تک نمونه و همبستگی پیرسون به منظور تحلیل داده ها استفاده شد. براساس نتایج پژوهش، کیفیت t آماری سبک زندگی زنان باردار مورد مطالعه، در حد متوسطی گزارش شد. همچنین مقایسهٔ شاخص های وزن، قد تفاوت که نشان از .(P≤0/ نشان داد (۰۵ NCHS معناداری را در مقطعی از بازه های زمانی پس از تولد در مقایسه با هنجار پایین تر بودن این هنجار در نوزادان ایرانی در قیاس با هنجار جهانی است. از طرف دیگر، دور سر نوزادان تفاوت با توجه به یافته های پژوهش در خصوص سبک زندگی و فعالیت بدنی متوسط در .(P=0/ معناداری را نشان نداد ۱۸ زنان باردار، احتمالاً توصیهٔ مادران به انجام فعالیت بدنی بیشتر و بهبود کیفیت تغذیهٔ آنها بتواند سبک زندگی آنها را ارتقا دهد و از این طریق بتوان شاخص های رشدی نوزادان را بهبود بخشید. **عنوان مرجع (reference title):** مطالعه تأثیر سبک زندگی مادران باردار بر شاخص‌های رشد جسمانی نوزادان پس از تولد **mT5:** بررسی ارتباط سبک زندگی مادران باردار بر شاخص های رشد جس **عنوان تولیدشده (معماری شماره ۲):** مطالعه تأثیر سبک زندگی مادران باردار بر شاخص‌های رشد جسمانی نوزادان پس از تولد

جدول ۵–۶ دقت مدل پیشنهادی ۲ به همراه دقت مدل عنوان‌ساز mT5

مدل	ROUGE-L	ROUGE-2	ROUGE-1
mT5	۴۱.۷۲	۲۷.۴۶	۴۵.۰۳
پیشنهادی شماره ۲ (5*mT5+selector)	**۴۲.۵۳**	**۲۸.۲۷**	**۴۶.۳۲**

۵-۳-۳ مقایسه‌ها

در جدول مقایسه‌ای از چند مدل آورده شده. علاوه بر مدل‌هایی پیشنهاد شده که در بخش قبل صحبت شد. در اینجا دقت یک مدل mT5-small تنظیم شده هم آورده شده است. این مدل نسبت به نسخه‌ی base تعداد کمتری پارامتر دارد و در نتیجه حجم کمتری دارد. همچنین یک مدل mT5-base را با داده‌های pn-summary که در مقاله [49] معرفی شده است به اندازه سه دوره آموزش دادیم. این مجموعه داده از سایت‌های خبری فارسی برای عمل خلاصه‌سازی جمع‌آوری شده ما برای کار خودمان تنها متن و عناوین های متناظر را جدا کردیم و مدل mT5-base را از ابتدا برای ۳ دور آموزش دادیم.

جدول ۵–۷ مقایسه مدل‌ها با یکدیگر توسط دقت هر یک با دادگان مختلف

مدل‌ها	دقت (ROUGH-1)	دادگان
mT5-base	۴۵.۰۳	پیشنهادی این پایان‌نامه
	۴۴.۶۹	pn-summary (خبری)
پیشنهادی شماره ۱ (mT5+denoiser)	۴۵.۹۵	پیشنهادی این پایان‌نامه
	۴۵.۲۵	pn-summary (خبری)
پیشنهادی شماره ۲ (5*mT5+selector)	**۴۶.۳۲**	پیشنهادی این پایان‌نامه
	۴۶.۲۰	pn-summary (خبری)
mT5-small	۳۹.۹۱	پیشنهادی این پایان‌نامه
mT5-base (آموزش داده شده با دادگان pn-summary)	۳۲.۹۲	پیشنهادی این پایان‌نامه
مدل پیشنهادی مرجع [51] (LSTM)	۴۲.۶	پیشنهادی مرجع [51]
[انگلیسی] مدل مرجع پیشنهادی مرجع [2]	۳۷.۲۷	پیشنهادی مرجع [2]

برای مشاهده نمونه عنوان‌های تولید شده توسط مدل‌های پیشنهادی می‌توانید به پیوست یک مراجعه کنید. همچنین برای دسترسی به مدل آموزش داده شده می‌توانید از لینک آدرس[1] در پاورقی، مدل را دریافت کنید یا به صورت آنلاین از آن استفاده کنید و خروجی بگیرید.

۵-۴ نتیجه‌گیری

در این بخش چند مدل و معماری مختلف برای پیاده‌سازی عنوان‌ساز فارسی چکیده‌ای ارائه شد. به دلیل اینکه برای اولین بار برای زبان فارسی انجام می‌شود، نمی‌توان مقایسه‌ای با کارهای دیگر انجام داد. با توجه به جدول ۵–۷ که چند مدل با دقت‌هایشان ارائه شده است، مدل پیشنهادی شماره ۲ (5*mT5_selector) بهترین نتیجه را با معیار ROUGE1 برای عمل عنوان‌سازی فارسی بدست آورده است. اما در صورتی که سرعت بیشتر مد نظر باشد، مدل‌های mT5-base و پیشنهادی اول (mT5+denoiser) پیشنهاد می‌شود.

[1] https://huggingface.co/Erfan/mT5-base_Farsi_Title_Generator

فصل ۶ جمع‌بندی و پیشنهادها

۶-۱ جمع‌بندی و نتیجه‌گیری

در این پایان‌نامه هدف پیاده‌سازی یک عنوان‌ساز بود تا با رویکرد چکیده‌ای بتواند برای یک متن کوتاه (مثل چکیده‌ی یک مقاله) عنوان مناسب تولید کند. برای شروع اینکار، نیاز به دادگانی که موضوع‌های مختلفی را در بر بگیرد نیاز بود تا این کار به خوبی انجام شود. برای اینکار، ابتدا از سایت‌های ژورنال‌های مقالات فراداده‌های مقالات بارگذاری شد که شامل چکیده‌ی مقالات و عنوان متناظر بودند. بعد از جمع‌آوری داده‌ها، یک مدل از پیش آموزش‌داده‌شده mT5-base را با داده‌های موجود آموزش دادیم. این مدل به تنهایی خروجی نویزدار تولید می‌کرد؛ به عنوان مثال قطعه‌های نامربوط در خروجی دیده می‌شود. برای غلبه بر این مشکل دو راهکار پیشنهاد کردیم. راهکار اول استفاده از یک ماژول نویززدا در ادامه‌ی مدل اصلی بود تا عنوان تولید شده را بهبود دهد. با اینکار، دقت مدل به اندازه یک واحد ROUGE1 افزایش پیدا کرد. در راهکار دوم پیشنهادی، تولید چند عنوان با مدل‌های آموزش دیده شده توسط دوره‌های مختلف است که در نهایت با انتخاب یکی با استفاده از معیار شباهت، آن عنوانی که شباهت بیشتری با متن مرجع دارد انتخاب می‌شود. با پیاده شدی این معماری دقت مدل کمتر از دو واحد ROUGE1 افزایش پیدا کرد. در نتیجه هر دو معماری پیشنهاد شده، دیگر قطعه‌های نامربوط و بی‌معنی بسیار کمتر در خورجی ظاهر می‌شود.

۶-۲ پیشنهادها برای ادامه فعالیت

برای ادامه کار پیشنهاد می‌شود که کارهای زیر نظر گرفته شود:

- می‌توان برای پیاده‌سازی خلاصه‌ساز چیکده‌ای فارسی، از مقاله‌های فارسی موجود به عنوان داده استفاده کرد. در این صورت، با پوشش موضوعی متنوع‌تر خلاصه‌ساز مناسبی را می‌شود آموزش داد. برای آموزش مدل، به بدنه اصلی مقاله به عنوان متن مرجع و چکیده‌ی متناظر به عنوان

خلاصه‌ی متن مرجع نیاز است. برای استخراج این داده‌ها نیاز به یک مدل شناخت کاراکترها و کلمات داریم تا فایل‌های pdf مقالات را به رشته‌های قابل محاسبه برای کامپیوتر تبدیل کند.

- برای ماژول نویززدا در ادامه مدل اصلی عنوان‌ساز می‌توان با تکنیک‌های خودناظر مدلی رو آموزش داد تا نویز را برطرف کند. به این صورت که، یک مدل از پیش‌آموزش داده شده‌ی مناسب را دوباره با عمل خودناظر جایگذاری[1] تنظیم می‌کنیم. برای این کار، متن مرجع و عنوان را در کنار هم گذاشته و به صورت رندم کلمه‌هایی از عنوان را با قطعه‌ی جای‌خالی پر می‌کنیم. سپس این رشته را به مدل ورودی می‌دهیم و مدل را آموزش می‌دهیم تا جاهای خالی را با کلمات مناسب پر کند.
- نیاز به راهکاری برای سبک کردن مدل‌ها و سرعت بیشتر است.

[1] Masking

پیوست

جدول نمونه خروجی‌های تولید شده

#	نمونه‌ها
۱	**متن چکیده:** میکروتوربین اغلب به سبب عواملی چ ون تغییرات مشخصات محیطی (فشـار، دمـا و ...) ، تغییـر میـزان تـوان مـورد نیـاز و راه اندازی سیستم تا رسیدن بـه بـار نـامی در خـارج از نقطـهٔ طراحی خود کار می کند . بنـابراین، بررسـی حالـت خـارج از نقطه طراحی میکروتوربین بسیار مفید و ضروری می باشد . در این مقاله، سیکل میکروتو ربین ۲۰۰ کیلـوواتی کپسـتون در خارج از نقطه طراحی مورد بررسی قرار می گیرد . بـرای ایـن منظور نیاز به منحنی مشخصه اجـزای مختلـف میکروتـوربین مذکور میباشد . در مطالعه حاضر، با توجه به ایـن نمودارهـا که از نتایج آزمایشگاهی بدست آمده اند، روابطی بین دبی هوا، نسبت فشار، دور و دمای ورودی توربین بدسـت مـی آیـد . بـا توجه به روابط استخراج شده و مدلسـازی انجـام شـده بـرای حالت طراحی، مدلسازی اجزای مختلـف در خـارج از نقطـه طراحی صورت می گیرد . باتوجه به این مدلسازی، تاثیر تغییـر دمای هوای ورودی نسبت بـه دمـای طراحـی و تغییـر تـوان نسبت به توا ن نامی در دو حالت دور ثابت و دور متغیر مـورد بررسی و مقایسه قرار میگیرد . نتایج نشان میدهد که با افزایش دمای هوای ورودی کمپرسور از ۱۵°C به ۴۵°C ، توان تولیدی حدود ٪۱۸ و بـازدهی حـدود ٪۲ نسبت به حالت طراحی کاهش می یابد . همچنـین، در ٪۷۰ توان نامی، بازدهی سیکل ٪۲۵ در حالت دور ثابت نسبت به حالت طراحی کاهش می یابد، در حالی کـه ایـن کـاهش در دور متغیر ٪۴ میباشد.
	عنوان مرجع : تحلیل عملکرد میکروتوربین ۲۰۰ کیلوواتی درخارج از نقطه طراحی
	معماری شماره ۱: بررسی مدلسازی اجزای مختلف میکروتوربین در خارج از نقطه
	معماری شماره ۲: بررسی حالت خارج از نقطه طراحی میکروتوربین کپستون
۲	**متن چکیده:** قلمرو حکومت عقل بر رد و قبول امور آیا ادیان را نیز در بر می گیرد؟ آیا ایمان حاصل برهان و استدلال است یا منشاء دیگری دارد؟ اگر در ایمان به برهان عقلی نیازی نیست پس منشاء آن چیست؟ در این مقاله پس از بحثی اجمالی در این مقولات به بررسی و نقد نظرات ابن جوزی پرداخته شده است. ابن جوزی – واعظ ، مفسر و فقیه حنبلی – با تأکید بر لزوم عقل و استدلال برای وصول به اعتقاد دینی ، کتاب تلبیس ابلیس را آغاز می کند، اما پس از نکوهش و سرزنش فیلسوفان و ناتوان از توفیق در مناظره با آنان خواننده کتاب خویش را میان تقلید – که عقل در آن راه ندارد – و استقلال عقلی – که ترک جمود بر ظواهر است – سردرگم رها می سازد.
	عنوان مرجع:

	عقل و استدلال در تفکر ابن جوزی
	معماری شماره ۱: نقد نظرات ابن جوزی
	معماری شماره ۲: بررسی و نقد نظرات ابن جوزی در باب عقل و استدلال
۳	**متن چکیده:** موج اسلام گرایی در آسیای مرکزی و همجواری این منطقه با ایالت مسلمان نشین و خود مختار سین‌کیانگ چین، خطر جدایی‌طلبی در این استان را تشدید می‌کند. شواهد زیادی وجود دارد که این یک تهدید واقعی است و ناشی از بزرگ نمایی چینی‌ها جهت سرکوب هر چه بیشتر معترضان مسلمان در سین‌کیانگ نمی‌باشد. سؤال اصلی مقاله این است که مهم‌ترین دغدغه امنیتی و سیاسی چین در آسیای مرکزی که منجر به افزایش روابط چین با این کشورها از ۲۰۰۵-۱۹۹۱ شده، چیست؟ فرضیه این مقاله آن است که مهم‌ترین دغدغه امنیتی و سیاسی چین در آسیای مرکزی هراس از بنیاد گرایی، افراط‌گرایی اسلامی و تروریسم است.
	عنوان مرجع: ملاحظات امنیتی- سیاسی چین در آسیای مرکزی (۲۰۰۵- ۱۹۹۱)
	معماری شماره ۱: امنیت و سیاست چین در آسیای مرکزی
	معماری شمار۲: بررسی خطر جدایی طلبی در آسیای مرکزی
۴	**متن چکیده:** بهبود کیفیت خط مشی های عمومی از جمله دغدغه ها و چالش های فراروی حوزه های مختلف عمومی همه نظام های سیاسی است. این کیفیت، در خلاء بهبود پیدا نمی کند، بلکه مستلزم داشتن راهکارها و ساز و کارهای مختلفی می باشد. هدف این پژوهش واکاوی و شناسایی ساز و کارهای دانشی بهبود کیفیت خط مشی های عمومی در ایران است. پژوهش در دو مرحله انجام شده است: در مرحله نخست (کیفی) ساز و کارها، احصاء شده و در مرحله دوم ضمن رتبه بندی اهمیت ساز و کار ها، فرضیه های توصیفی در مورد ساز و کارها در معرض آزمون قرار داده شد. نتایج نشان می دهد که ساز و کارهای دانشی برای بهبود خط مشی های عمومی به ترتیب اهمیت عبارتند از:کاربردی سازی دانش خط مشی گذاری عمومی در نظام خط مشی گذاری عمومی کشور، اشاعه دانش خط مشی گذاری عمومی در کشور و شکل دهی مبانی فلسفی و نظری بومی خط مشی گذاران عمومی.
	عنوان مرجع: واکاوی ساز و کارهای دانشی بهبود کیفیت خط‌مشی‌های عمومی در ایران: پژوهش ترکیبی
	معماری شماره ۱: واکاوی و شناسایی ساز و کارهای دانشی بهبود کیفیت خط مشی
	معماری شماره ۲: واکاوی و شناسایی ساز و کارهای دانشی بهبود کیفیت خط مشی

۵	**متن چکیده:** نقد تکوینی"" مبتنی بر مطالعهٔ فرآیند شکل گیری آثار بر اساس ""پیشامتن""ها یعنی بقایای مادی برجامانده از طرح ریزی های اولیه منتهی به شکل نهایی آثار است. این شیوهٔ نقد که در دههٔ ۷۰ قرن بیستم م و ابتدا در ادبیات پدید آمد و به تدریج در هنرهای دیگر، مانند نقاشی نیز به کار گرفته شد، امروزه در محافل هنری و مجامع علمی، علاقمندان فراوانی یافته و درحال پیشرفت است، اما هنوز درعرصهٔ عکاسی به کاربرده نشده است. گاه در نقد و بررسی عکس ها، مواجه با نوعی از شرح مراحل تولید عکس ها و یا مسائل پیرامونی خلق آنها هستیم، اما این گزارش ها، مبتنی بر اصول و روش نقد تکوینی نبوده و اهداف آن را محقق نمی سازند. در این مقاله، ابتدا پیدایش و روش نقد تکوینی و عناصر مربوط به آن، مانند پیرامتن، پیش متن و پیشامتن، بصورت اجمالی معرفی شده است. پس از آن با نگاه به تجارب منتقدان تکوینی در عرصه های دیگر و بررسی فرآیند های خلق عملی عکس، عنصر محوری نقد تکونی یعنی پیشامتن، در عکاسی، تعریف شده و امکان و شیوهٔ کاربرد نقد تکوینی در عکاسی مورد بحث قرار گرفته است. در پایان نیز به عنوان مثال، عکس ""پسرک با نارنجک دستی اسباب بازی"" اثر دایان آربُس، تحلیل شده است
	عنوان مرجع: نقد تکوینی عکس (بررسی امکان و شیوهٔ کاربرد نقد تکوینی در عکاسی)
	معماری شماره ۱: نقد تکوینی عکس
	معماری شماره ۲: نقد تکوینی عکس
۶	**متن چکیده:** حقیق پیشرو کنکاشی است در باب شکل گیری مکتب نئورئالیسم در ادبیات ایتالیا و پیشگامان آن، هم‌زمان فرونشستن شعله‌های آتش جنگ جهانی دوم در اروپا، موجی از تحولات و تغییرات تمام ابعاد زندگی مردم را در بر گرفت که به صورت یک تطابق سریع و جهانی از تمدن گسترش یافت. نئورئالیسم سعی در نمایاندن رویدادهای اجتماعی و سیاسی دارد و تمام لایه‌ها و طبقات جامعهٔ معاصر را از نظر تأثیر جنگ، فاشیسم و ظهور نیروهای سیاسی نوین بر آنها ترسیم می‌نماید. این تحقیق سعی در معرفی مشهورترین نویسندگان سبک نئورئالیسم ادبیات ایتالیا دارد؛ من‌جمله: چه زاره پاورزه Cesare Pavese اِلیو ویتّورینی Elio Vittorini واسکو پراتولینی Vasco Pratolini فرانچسکو یووینه Francesco Jovine لئوناردو شاشا Leonardo Sciascia کارلو بِرناری Carlo Bernari جوزپّه برتو Giuseppe Berto جوزپّه دسی Giuseppe Dessi اِنریکو اِمانوئلّی Enrico Emanuelli جانّا مانزینی Gianna Manzini ناتالیا گینزبورگ Natalia Ginsburg دُمنیکو رِآ Domenico Rea جووانی تستوری Giovanni Testori آلبرتو پینکرله Alberto Pincherle
	عنوان مرجع: ادبیات ایتالیا بعد از جنگ جهانی دوم
	معماری شماره ۱: نئورئالیسم در ادبیات ایتالیا

	معماری شماره ۲: شکل گیری مکتب نئورئالیسم در ادبیات ایتالیا
۷	**متن چکیده:** انتقال معتبر و غیرمجاز وجوه یکی از مهم ترین خطرات در بانکداری الکترونیکی محسوب می شود. اگر دستور پرداخت معتبر و غیرمجازی صادر شود و بانک از غیرمجاز بودن دستور پرداخت اطلاعی نداشته باشد و دستور پرداخت به وسیله روش های امنیتی مورد توافق با مشتری شناسایی و تصدیق شود، بانک دستور پرداخت را معتبر تشخیص می دهد و نسبت به انتقال وجه اقدام خواهد کرد. در چنین شرایطی آیا بانک یا شخص دیگری ضمان و مسئولیت مدنی دارند؟ با توجه به اینکه دستور پرداخت به وسیله چه ابزار پرداختی صادر شده و اینکه صدور چنین دستور پرداختی ناشی از قصور مشتری یا بانک بوده است یا اینکه هیچ یک از بانک و مشتری تقصیری مرتکب نشده باشد، موجب می شود ضمان و مسئولیت جبران خسارت بر عهده بانک یا مشتری یا هر دوی آنها قرار گیرد. در این مقاله، مفهوم انتقال غیرمجاز و انتقال معتبر وجوه، انواع دستور پرداخت از حیث جواز و اعتبار و مبانی مسئولیت مدنی و آثار ناشی از نقض تعهد بر اساس حقوق ایران و قانون متحدالشکل تجاری آمریکا بررسی می شود. به نظر نگارندگان در پرداخت معتبر و غیرمجاز اصل بر ضمان و مسئولیت مدنی بانک است؛ مگر آنکه بانک تقصیر و انتساب این امر به مشتری را ثابت کند. البته مشتری باید غیرمجاز بودن را ثابت کند.
	عنوان مرجع: ضمان و مسئولیت مدنی در انتقال الکترونیکی معتبر و غیرمجاز وجوه
	معماری شماره ۱: بررسی ضمان و مسئولیت مدنی بانک در انتقال غیرمجاز
	معماری شماره ۲: بررسی مفهوم انتقال غیرمجاز و انتقال معتبر وجوه در بانک
۸	متن چکیده: گروگان‌گیری جرمی است که با هدف واداشتن شخص ثالث به ارتکاب یا عدم ارتکاب رفتاری، از دیرباز مورد توجه مجرمان بوده است. ضرورت مقابله با این جرم به دلیل آثار و تبعات داخلی و بین‌المللی آن، باعث توجه بیش‌ازپیش نظام‌های حقوق کیفری به مقابله با آن در سطح ملی و بین‌المللی شده است. جرم‌انگاری این موضوع در اساسنامه‌ی دیوان کیفری بین‌المللی (۱۹۹۸)، به‌عنوان یکی از مصادیق جنایات جنگی نشانگر عمق نگرانی جامعه‌ی بین‌الملل نسبت به این جرم است. کنوانسیون مقابله با گروگان‌گیری (۱۹۷۹)، به‌عنوان مهم‌ترین سند اختصاصی تدوین‌شده در خصوص این جرم، کشورهای عضو را موظف به وضع ضمانت اجرای کیفری برای مرتکبین این جرم نموده است. در حقوق ایران، به‌رغم تصویب این کنوانسیون در مجلس شورای اسلامی، هنوز ضمانت اجرای کیفری خاصی برای مرتکبان این جرم پیش‌بینی نشده است. ازاین‌رو در موارد ارتکاب این جرم، رویه‌ی یکسانی در جهت تعیین مجازات برای متهمان به ارتکاب جرم مذکور وجود ندارد. در این نوشتار با بررسی این کنوانسیون و سایر اسناد بین‌المللی و مقررات مرتبط در حقوق ایران، با تبیین ارکان تشکیل‌دهنده‌ی جرم مذکور، مجازات‌های قابل اِعمال در خصوص این جرم در نظام کیفری ایران مورد بررسی قرار گرفته است.

	عنوان مرجع: ضمانت اجرای کیفری گروگان گیری در حقوق ایران با نگاهی به اسناد بین المللی
	معماری شماره ۱: بررسی مجازات های قابل اعمال در خصوص جرم گروگان گیری
	معماری شماره ۲: بررسی جرم مقابله با گروگان گیری جرمی در نظام کیفری ایران
۹	**متن چکیده:** به منظور بررسی اثر روش‌های مختلف فراوری دانه گندم با منابع مختلف چربی جیره بر روی عملکرد پروار، خصوصیات لاشه و پروفیل اسیدهای چرب گوشت، ۲۸ راس گوساله نر هلشتاین با میانگین وزنی ۵۶±۲۹۶ کیلوگرم، به چهار جیره آزمایشی (۷ گوساله در هر تیمار) اختصاص داده شد. این آزمایش به صورت فاکتوریل (دو روش فراوری دانه گندم شامل ورقه ای کردن با بخار و گندم فراوری شده با فرمالدهید، و دو منبع چربی جیره شامل دانه سویای برشته شده و پودر چربی)رومی فت۱) و در قالب طرح کاملا تصادفی انجام شد. مدت انجام این آزمایش ۹۸ روز ، که ۱۴ روز آن به عنوان دوره عادت‌دهی در نظر گرفته شد. به طور افرادی ماده خشک مصرفی به صورت روزانه و افزایش وزن گوساله‌ها به صورت ماهانه اندازه‌گیری شد. به دنبال وزن کشی نهایی در روز ۸۵، جهت بررسی خصوصیات لاشه از هر تیمار ۳ گوساله کشتار گردید. اختلاف ماده خشک مصرفی، میانگین افزایش وزن روزانه و بازدهی غذایی در بین تیمارهای آزمایشی معنی دار نشد. هم چنین هیچ کدام از صفات مربوط به خصوصیات لاشه، تحت تاثیر جیره‌های آزمایش قرار نگرفت. مقدار C ۱۸:۲، C ۱۸:۳، C ۲۴:۰ و کل اسیدهای چرب غیر اشباع در گوشت حاصل از گوساله های مصرف کننده سویای برشته شده بیشتر بود ($p<۰/۰۱$). هم چنین مصرف دانه سویای برشته شده باعث افزایش مقدار C ۱۸:۰ و اسید لینولئیک کونژوگه شده در گوشت گوساله‌ها شد($p<۰/۰۲$). در مقابل مقدار C۱۶:۰ و کل اسید های چرب اشباع در گوشت حاصل ازگوساله‌های مصرف‌کننده پودر چربی بیشتر بود (به ترتیب $p<۰/۰۱$ و $p<۰/۰۵$). سایر اسیدهای چرب موجود در گوشت گوساله ها تحت تاثیر جیره های آزمایشی قرار نگرفت. نتایج حاصل از این مطالعه نشان می دهد که عملکرد پروار در گوساله‌های نر هلشتاین تحت تاثیر روش های مختلف فراوری دانه گندم و منابع چربی جیره قرار نگرفت. اما استفاده از دانه سویای برشته شده در جیره گوساله ها باعث بهبود پروفیل اسیدهای چرب راسته، از لحاظ اهمیت آن ها در تغذیه انسان شد.
	عنوان مرجع: اثر فراوری دانه گندم با منابع مختلف چربی جیره بر عملکرد و پروفیل اسیدهای چرب راسته گوساله‌های نر هلشتاین
	معماری شماره ۱: اثر روش‌های مختلف فراوری دانه گندم با منابع چربی جیره
	معماری شماره ۲: اثر روش های مختلف فراوری دانه گندم با منابع مختلف چربی
۱۰	**متن چکیده:**

	هدف از این مطالعه، شناسایی عوامل مؤثر بر تمایل به پرداخت شهروندان مشهد برای دمنوش‌های گیاهی بود. داده‌ها به روش پیمایشی و با تکمیل ۳۶۸ پرسشنامه به روش نمونه‌گیری تصادفی ساده از میان شهروندان مشهد، در بهار و تابستان سال ۱۳۹۶ جمع‌آوری شد. به‌منظور شناسایی عوامل مؤثر بر تمایل به پرداخت مصرف‌کنندگان، از روش ارزش‌گذاری مشروط و برآورد الگوی لاجیت ترتیبی استفاده شد. تحلیل داده‌ها نشان داد که ۵۰ درصد مصرف‌کنندگان، تمایل به پرداخت ۵ الی ۲۵ درصد هزینه بیشتر برای خرید دمنوش‌های گیاهی داشته‌اند. نتایج برآورد الگو نشان داد که درآمد خانوار، آگاهی از خواص درمانی، بسته‌بندی، نداشتن اسانس، نحوه عرضه، داشتن برچسب اطلاعات، جنسیت زن و سطح تحصیلات مصرف‌کنندگان، اثر مثبت و معنی‌دار و سن، بعد خانوار، کیفیت، فرآوری و قیمت دمنوش‌ها اثر منفی و معنی‌دار بر تمایل به پرداخت داشته‌اند. بنا بر نتایج، با استفاده از برچسب اطلاعات درباره محصول، بسته‌بندی مناسب و همچنین، آگاهی‌بخشی درباره خواص و نحوه مصرف محصول، می‌توان احتمال تمایل به پرداخت بیش از ۱۵ درصد هزینه اضافی برای دمنوش‌های گیاهی را به ترتیب به اندازه ۷۱، ۳۶ و ۴۲ درصد افزایش داد.
	عنوان مرجع: شناسایی عوامل تعیین‌کننده میزان تمایل به پرداخت شهروندان مشهد برای دمنوش‌های گیاهی
	معماری شماره ۱: شناسایی عوامل مؤثر بر تمایل به پرداخت شهروندان مشهد
	معماری شماره ۲: بررسی عوامل مؤثر بر تمایل به پرداخت شهروندان مشهد
۱۱	**متن چکیده:** در این نوشتار، انتقال الکترونیکی وجوه از دریچه نظریه انتقال تعهد بررسی‌شده است. ازاین‌روی، دستور پرداخت یا انتقال به‌مانند ایجاب قرارداد شمرده می‌شود. اگر انتقال وجه بخواهد در چهارچوب این نظریه بگنجد، باید شرایط پیدایش و نیز آثار حقوقی آن را داشته باشد. از نگاه شرایط عمل، طلب یا دین به‌عنوان موضوع قرارداد در بیشتر حالات موجود است و حتی اگر میان محیل (دستوردهنده) و محتال (ذی‌نفع پرداخت) رابطه دینی نبوده و دست جستن به حواله شدنی نباشد، دریچه استناد به انتقال طلب باز است. از دید قصد، چون دستوردهنده جابجایی پولی را می‌خواهد که نزد بانک دارد، به دلیل نقش نداشتن اراده بدهکار و نبود رابطه امانی، این خواسته را باید بر انتقال طلب بار نمود. از دریچه نقش‌آفرینی رضایت یا آگاهی بانک، اگر سرشت حقوقی بر پایه نظریه انتقال تعهد بررسی گردد، حق بانک برای نپذیرفتن دستور به گونه کامل توجیه‌پذیر است. البته، در انتقال طلب که رضایت بدهکار نقشی ندارد، بررسی که باتک انجام می‌دهد، پیرامون لوازم انتقال وجه است نه انشای قرارداد. با این همه ، چون نظریه انتقال تعهد پاسخ برخی از حالات انتقال وجه همانند انتقال به‌حساب خود را نمی‌دهد، مورد نکوهش است
	عنوان مرجع: دستور پرداخت در انتقال الکترونیکی وجه به مانند ایجاب انتقال تعهد
	معماری شماره ۱: انتقال الکترونیکی وجوه از دریچه نظریه انتقال تعهد
	معماری شماره ۲: بررسی انتقال الکترونیکی وجوه از دیدگاه نظریه انتقال تعهد

لغت‌نامه

مرتب بر اساس حروف الفبای فارسی

عنوان فارسی	عنوان انگلیسی
ابر پارامتر	Hyperparameter
اجرا در هر آزمایش	Executions per trial
استخراجی	Extractive
اعمال پایین‌دستی	Down-stream tasks
افزونگی	Redundancy
اندازه هسته	Kernel Size
انفجار شیب	Exploding Gradient
بازیابی اطلاعات	Information retrival
بدون‌ناظر	Unsupervised
برچسب‌گذاری قواعد معنایی	Semantic Role Labeling
بردار	Vector
پدینگ	Padding
پردازش زبان طبیعی	Natural Language Processing
پیش‌آموزش	Pre-trained
پیکره متون	Corpus
تابع بهینه سازی	Optimizer Function
تابع خطا	Loss Function
تابع فعال ساز	Activation Function
تابع فعال‌ساز سیگموئید	Sigmoid activation function
تانژانت هیپربولیک	Tangent Hyperbolic
تجزیه‌کننده	Parser
تحلیل احساسی	Sentiment analisys
تکنیک ماسک کردن	Masking
تماما متصل	Fully connected

Embedding	جاسازی
Word Embedding	جاسازی کلمات
Abstractive	چکیده‌ای
Feed-Forward	حرکت روبه‌جلو
Out of vocabulary	خارج از دایره واژگانی
Web crawling	خزش وب
Linearization	خطی‌سازی
Multi-document summarizer	خلاصه‌ساز چند متنی
Automatic Text Summarization	خلاصه‌ساز خودکار متون
Self-attention	خود توجه
Data-driven	داده‌محور
Sequential data	داده‌های دنباله‌ای
Dependency	درخت تجزیه وابستگی
Partial dependency trees	درخت‌های وابستگی جزئی
Training batch	دسته‌های آموزشی
Validation Accuracy	دقت اعتبار سنجی
Phrase sequence	دنبال عبارات
Sequence to Sequence	دنباله به دنباله
Epoch	دوره
String	رشته
Decoder	رمزگشا
Encoder	رمزنگار
Semantical error	روش‌های معنایی
Lexical chain	زنجیره واژگانی
Subpath	زیرمسیر
Predicate argument structure	ساختار نهاد گزاره
Recommender Systems	سیستم‌های پیشنهاددهنده
Dual-Path Network	شبکه دومسیره
Transformer networks	شبکه‌های انتقالی
Recurrent Neural Networks	شبکه‌های بازگشتی
Input Lenght	طول ورودی

Metadata	فراداده
Filter	فیلتر
Token	قطعه
Tokenization	قطعه‌بندی
Convolutional	کانالوشنی
Pytorch library	کتابخوانه پایتورچ
Pointer-generator switch	کلید تولید نشانه‌گر
Train step	گام آموزشی
Vanishing gradient	گرادیان محو شونده
Node	گره
Back-propagation alghorithms	الگوریتم‌های انتشار رو به عقب
Dictionary	لغت نامه
Support vector machine	ماشین بردار پشتیبان
Text-to-Text	متن به متن
Validation Set	مجموعه ارزیابی
Text-to-Text Transfer Transformer	مدل انتقالی متن به متن
Path	مسیر
Jaccard Criteria	معیار جاکارد
Attention mechanism	مکانیزم توجه
Vanishing Gradient	ناپدید شدن شیب
Named entity recognition	نامگذاری موجودیت‌ها
Dropout Rate	نرخ حذف تصادفی
Flag	نشانه
Represantation	نمایش
Variance	واریانس
Overlap	همپوشانی
Transfer Learning	یادگیری انتقالی
Supervised learning	یادگیری باناظر
Deep learning	یادگیری عمیق

مرتب بر اساس حروف الفبای انگلیسی

عنوان فارسی	عنوان انگلیسی
Abstractive	چکیده‌ای
Activation Function	تابع فعال ساز
Attention mechanism	مکانیزم توجه
Automatic Text Summarization	خلاصه‌ساز خودکار متون
Back-propagation alghorithms	الگوریتم‌های انتشار رو به عقب
Convolutional	کانالوشنی
Corpus	پیکره متون
Data-driven	داده‌محور
Decoder	رمزگشا
Deep learning	یادگیری عمیق
Dependency	درخت تجزیه وابستگی
Dictionary	لغت نامه
Down-stream tasks	اعمال پایین‌دستی
Dropout Rate	نرخ حذف تصادفی
Dual-Path Network	شبکه دومسیره
Embedding	جاسازی
Encoder	رمزنگار
Epoch	دوره
Executions per trial	اجرا در هر آزمایش
Exploding Gradient	انفجار شیب
Extractive	استخراجی
Feed-Forward	حرکت روبه‌جلو
Filter	فیلتر
Flag	نشانه
Fully connected	تماما متصل
Hyperparameter	ابر پارامتر
Information retrival	بازیابی اطلاعات
Input Lenght	طول ورودی
Jaccard Criteria	معیار جاکارد

Kernel Size	اندازه هسته
Lexical chain	زنجیره واژگانی
Linearization	خطی‌سازی
Loss Function	تابع خطا
Masking	تکنیک ماسک کردن
Metadata	فراداده
Multi-document summarizer	خلاصه‌ساز چند متنی
Named entity recognition	نامگذاری موجودیت‌ها
Natural Language Processing	پردازش زبان طبیعی
Node	گره
Optimizer Function	تابع بهینه سازی
Out of vocabulary	خارج از دایره واژگانی
Overlap	همپوشانی
Padding	پدینگ
Parser	تجزیه‌کننده
Partial dependency trees	درخت‌های وابستگی جزئی
Path	مسیر
Phrase sequence	دنبال عبارات
Pointer-generator switch	کلید تولید نشانه‌گر
Predicate argument structure	ساختار نهاد گزاره
Pre-trained	پیش‌آموزش
Pytorch library	کتابخوانه پایتورچ
Recommender Systems	سیستم‌های پیشنهاددهنده
Recurrent Neural Networks	شبکه‌های بازگشتی
Redundancy	افزونگی
Represantation	نمایش
Self-attention	خود توجه
Semantic Role Labeling	برچسب‌گذاری قواعد معنایی
Semantical error	روش‌های معنایی
Sentiment analisys	تحلیل احساسی
Sequence to Sequence	دنباله به دنباله

Sequential data	داده‌های دنباله‌ای
Sigmoid activation function	تابع فعال‌ساز سیگموئید
String	رشته
Subpath	زیرمسیر
Supervised learning	یادگیری باناظر
Support vector machine	ماشین بردار پشتیبان
Tangent Hyperbolic	تانژانت هیپربولیک
Text-to-Text	متن به متن
Text-to-Text Transfer Transformer	مدل انتقالی متن به متن
Token	قطعه
Tokenization	قطعه‌بندی
Train step	گام آموزشی
Training batch	دسته‌های آموزشی
Transfer Learning	یادگیری انتقالی
Transformer networks	شبکه‌های انتقالی
Unsupervised	بدون‌ناظر
Validation Accuracy	دقت اعتبار سنجی
Validation Set	مجموعه ارزیابی
Vanishing gradient	گرادیان محو شونده
Vanishing Gradient	ناپدید شدن شیب
Variance	واریانس
Vector	بردار
Vocabulary	واژگان
Web crawling	خزش وب
Word Embedding	جاسازی کلمات

مراجع

[1] N. Nazari and M. Mahdavi, “A survey on Automatic Text Summarization,” *J. AI Data Min.*, vol. 0, no. 0, pp. 121–135, 2018, doi: 10.22044/jadm.2018.6139.1726.

[2] E. Çano and O. Bojar, “Two huge title and keyword generation corpora of research articles,” *Lr. 2020 - 12th Int. Conf. Lang. Resour. Eval. Conf. Proc.*, pp. 6663–6671, 2020.

[3] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, “Automatic text summarization: A comprehensive survey,” *Expert Syst. Appl.*, vol. 165, p. 113679, 2021, doi: 10.1016/j.eswa.2020.113679.

[4] S. Gupta and S. K. Gupta, “Abstractive summarization: An overview of the state of the art,” *Expert Syst. Appl.*, vol. 121, no. 2018, pp. 49–65, 2019, doi: 10.1016/j.eswa.2018.12.011.

[5] A. Khan, N. Salim, and Y. Jaya Kumar, “A framework for multi-document abstractive summarization based on semantic role labelling,” *Appl. Soft Comput. J.*, vol. 30, pp. 737–747, 2015, doi: 10.1016/j.asoc.2015.01.070.

[6] L. J. Kurisinkel, Y. Zhang, and V. Varma, “Abstractive Multi-document Summarization by Partial Tree Extraction, Recombination and Linearization,” *8th Int. Jt. Conf. Nat. Lang. Process.*, pp. 812–821, 2017.

[7] A. See, P. J. Liu, and C. D. Manning, “Get to the point: Summarization with pointer-generator networks,” *ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 1, pp. 1073–1083, 2017, doi: 10.18653/v1/P17-1099.

[8] A. A. Syed, F. L. Gaol, and T. Matsuo, “A Survey of the State-of-the-Art Models in Neural Abstractive Text Summarization,” *IEEE Access*, vol. 9, pp. 13248–13265, 2021, doi: 10.1109/ACCESS.2021.3052783.

[9] A. Magueresse, V. Carles, and E. Heetderks, “Low-resource Languages: A Review of Past Work and Future Challenges,” *Computation and Language (cs.CL)* ,2020, [Online]. Available: http://arxiv.org/abs/2006.07264.

[10] R. Nallapati, B. Zhou, C. N. dos Santos, C. Gulcehre, and B. Xiang, “Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond,” *CoNLL 2016 - 20th SIGNLL Conf. Comput. Nat. Lang. Learn. Proc.*, pp. 280–290, Feb. 2016, Accessed: Dec. 02, 2021. [Online]. Available: http://arxiv.org/abs/1602.06023.

[11] A. Vaswani *et al.*, “Attention Is All You Need,” In *Advances in neural information processing systems* ,pp. 5998-6008 ,Jun. 2017, Accessed: Sep. 29, 2021. [Online]. Available: https://arxiv.org/abs/1706.03762.

[12] C. Raffel *et al.*, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *J. Mach. Learn. Res.*, vol. 21, pp. 1–67, 2020.

[13] A. R. Openai, K. N. Openai, T. S. Openai, and I. S. Openai, “Improving Language Understanding by Generative Pre-Training,” Accessed: Oct. 20, 2021. [Online]. Available: https://gluebenchmark.com/leaderboard.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 4171–4186, Oct. 2018, Accessed: Oct. 20, 2021. [Online]. Available: https://arxiv.org/abs/1810.04805v2.

[15] C. J. Van Rijsbergen, “INFORMATION RETRIEVAL: THEORY AND PRACTICE,” In *Proceedings of the Joint IBM/University of Newcastle upon Tyne Seminar on Data Base Systems*, pp. 1-14, 1979.

[16] G. Salton, “Automatic processing of foreign language documents,” *J. Am. Soc. Inf. Sci.*, vol. 21, no. 3, pp. 187–194, May 1970, doi: 10.1002/ASI.4630210305.

[17] G. Salton, “Syntactic approaches to automatic book indexing,” Proceedings of the 26th annual meeting on Association for Computational Linguistics, pp. 204–210, 1988, doi: 10.3115/982023.982048.

[18] M. Joshi, H. Wang, and S. McClean, “Dense Semantic Graph and Its Application in Single Document Summarisation BT - Emerging Ideas on Information Filtering and Retrieval: DART 2013: Revised and Invited Papers,” C. Lai, A. Giuliani, and G. Semeraro, Eds. Cham: Springer International Publishing, 2018, pp. 55–67.

[19] A. Nenkova and K. McKeown, “A Survey of Text Summarization Techniques BT - Mining Text Data,” C. C. Aggarwal and C. Zhai, Eds. Boston, MA: Springer US, 2012, pp. 43–76.

[20] R. A. García-Hernández and Y. Ledeneva, “Word sequence models for single text summarization,” *Proc. 2nd Int. Conf. Adv. Comput. Interact. ACHI 2009*, pp. 44–48, 2009, doi: 10.1109/ACHI.2009.58.

[21] J. Ramos, “Using TF-IDF to Determine Word Relevance in Document Queries,” *Proc. first Instr. Conf. Mach. Learn.*, vol. 242, no. 1, pp. 29–48, 2003.

[22] R. A. García-Hernández, R. Montiel, Y. Ledeneva, E. Rendón, A. Gelbukh, and R. Cruz, “Text summarization by sentence extraction using unsupervised learning,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5317 LNAI, pp. 133–143, 2008, doi: 10.1007/978-3-540-88636-512.

[23] M. Afsharizadeh, H. Ebrahimpour-Komleh, and A. Bagheri, “Query-oriented text summarization using sentence extraction technique,” *2018 4th Int. Conf. Web Res. ICWR 2018*, pp. 128–132, 2018, doi: 10.1109/ICWR.2018.8387248.

[24] R. Barzilay and M. Elhadad, “Using Lexical Chains for Text Summerization,” *Proc. ACL Work. Intell. Scalable Text Summ.*, pp. 10–17, 1997.

[25] C. Fellbaum, “WordNet: An electronic lexical database and some of its applications.” MIT press Cambridge, 1998.

[26] M. Ramezani and M.-R. Feizi-Derakhshi, “Ontology-based automatic text summarization using FarsNet,” *Adv. Comput. Sci. an Int. J.*, vol. 4, no. 2, pp. 88–96, 2015.۱

[27] M. Shamsfard, “Towards semi automatic construction of a lexical ontology for Persian,” *Proc. 6th Int. Conf. Lang. Resour. Eval. Lr. 2008*, pp. 2629–2633, 2008.

[28] B. Mohit, “Named entity recognition,” in *Natural language processing of semitic languages*, Springer, 2014, pp. 221–245.

[29] N. Moratanch and S. Chitrakala, “A survey on extractive text summarization,” in *2017 international conference on computer, communication and signal processing (ICCCSP)*, 2017, pp. 1–6.

[30] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, “Automatic text summarization: A comprehensive survey,” *Expert Syst. Appl.*, vol. 165, p. 113679, 2021.

[31] J. Ramos, “Using tf-idf to determine word relevance in document queries,” in *Proceedings of the first instructional conference on machine learning*, 2003, vol. 242, no. 1, pp. 29–48.

[32] M. E. Khademi and M. Fakhredanesh, “Persian Automatic Text Summarization Based on Named Entity Recognition,” *Iran. J. Sci. Technol. - Trans. Electr. Eng.*, vol. 2, 2020, doi: 10.1007/s40998-020-00352-2.

[33] A. AleAhmad, H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian, “Hamshahri: A standard Persian text collection,” *Knowledge-Based Syst.*, vol. 22, no. 5, pp. 382–387, 2009.

[34] A. Tandel, B. Modi, P. Gupta, S. Wagle, and S. Khedkar, “Multi-document text summarization-a survey,” in *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, 2016, pp. 331–334.

[35] J. Hou, C. Qiu, Y. Shen, H. Li, and X. Bao, “Engineering of Saccharomyces cerevisiae for the efficient co-utilization of glucose and xylose.,” *FEMS Yeast Res.*, vol. 17, no. 4.

[36] K. Ganesan, C. X. Zhai, and J. Han, “Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions,” *Coling 2010 - 23rd Int. Conf. Comput. Linguist. Proc. Conf.*, vol. 2, pp. 340–348, 2010.

[37] J. M. Czum, “Dive Into Deep Learning,” *J. Am. Coll. Radiol.*, vol. 17, no. 5, pp. 637–638, 2020, doi: 10.1016/j.jacr.2020.02.005.

[38] L. Hou, P. Hu, and C. Bei, “Abstractive document summarization via neural model with joint attention,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10619 LNAI, pp. 329–338, 2018, doi: 10.1007/978-3-319-73618-1_28.

[39] V. Gupta, N. Bansal, and A. Sharma, “Text summarization for big data: A comprehensive survey,” in *International Conference on Innovative Computing and Communications*, 2019, pp. 503–516.

[40] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv Prepr. arXiv1409.0473*, 2014.

[41] R. Paulus, C. Xiong, and R. Socher, “A deep reinforced model for abstractive summarization,” *6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc.*, no. i, pp. 1–12, 2018.

[42] S. Song, H. Huang, and T. Ruan, “Abstractive text summarization using LSTM-CNN based deep learning,” *Multimed. Tools Appl.*, vol. 78, no. 1, pp. 857–875, 2019, doi: 10.1007/s11042-018-5749-3.

[43] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, “PEGASUS: Pre-Training with extracted gap-sentences for abstractive summarization,” *37th Int. Conf. Mach. Learn. ICML 2020*, vol. PartF16814, pp. 11265–11276, 2020.

[44] C. Raffel *et al.*, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” *J. Mach. Learn. Res.*, vol. 21, pp. 1–67, Oct. 2019, Accessed: Sep. 19, 2021. [Online]. Available: https://arxiv.org/abs/1910.10683v3.

[45] E. Zolotareva, T. M. Tashu, and T. Horváth, “Abstractive text summarization using transfer learning,” *CEUR Workshop Proc.*, vol. 2718, pp. 75–80, 2020.

[46] V. Chen and E. T. Montaño, “An Examination of the CNN / DailyMail Neural Summarization Task,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 2358-2367. 2017.

[47] M. Shamsfard, “Challenges and Opportunities in Processing Low Resource Languages: A study on Persian,” pp. 291–295, Accessed: Oct. 20, 2021. [Online]. Available: https://dumps.wikimedia.org/fawiki/.

[48] L. Xue *et al.*, “mt5: A massively multilingual pre-trained text-to-text transformer,” *arXiv Prepr. arXiv2010.11934*, 2020.

[49] M. Farahani, M. Gharachorloo, and M. Manthouri, “Leveraging ParsBERT and Pretrained mT5 for Persian Abstractive Text Summarization,” *26th Int. Comput. Conf. Comput. Soc. Iran, CSICC 2021*, 2021, doi: 10.1109/CSICC52343.2021.9420563.

[50] M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri, “ParsBERT: Transformer-based model for Persian language understanding,” *Neural Process. Lett.*, vol. 53, no. 6, pp. 3831–3847, 2021.

[51] M. Mohseni and H. Faili, “Title Generation and Keyphrase Extraction from Persian Scientific Texts,” *2020 25th Int. Comput. Conf. Comput. Soc. Iran, CSICC 2020*, 2020, doi: 10.1109/CSICC49403.2020.9050113.

[52] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, “Opennmt: Open-source toolkit for neural machine translation,” *arXiv Prepr. arXiv1701.02810*, 2017.

[53] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Text summarization branches out*, pp. 74–81, 2004.

[54] E. M. B. Nagoudi, M. Abdul-Mageed, and A. Elmadany, “AraT5: Text-to-Text Transformers for Arabic Language Understanding and Generation,” *arXiv preprint arXiv:2109.12068*, 2021, [Online]. Available: http://arxiv.org/abs/2109.12068.

Abstract

Persian is a low-resource language and thus abstractive-summarization is a considerable challenge with the lack of an efficient dataset. We gathered a dataset suitable for fine-tuning a pretrained transformer model, then employed the mt5 pre-trained model to train an abstractive title-generation model on our introduced dataset. The model achieved the Rouge-1 score of 46.32 with coherent outputs. The results are competitive with state-of-the-art similar English works**.**

Keywords: ***Summarization, Deep learning, Deep Neural networks, Transformers, T5, MT5, Farsi Abstractive text Summarization,*** **Farsi Title-Generation**

Shahrood University of Thchnology

Faculty of Computer Engineering and Information Technology
M.Sc. Thesis in Artificial Intelligence Engineering

Summarization of short Persian documents using data-driven solutions

By: Erfan Jalili Jalal

Supervisors:
Dr. Marziea Rahimi

Advisors:
Dr. Morteza Zahedi

January 2021