Persian Text Summarization using Sparse Coding with Neural Text Representation

Ramin Fatourechi
MSc Artificial Intelligence; Department of Computer Engineering; Amirkabir University of Technology; Tehran, Iran;
Email: r.fatourechi@aut.ac.ir

Saeedeh Momtazi*
PhD Artificial Intelligence; Assistant Professor; Department of Computer Engineering; Amirkabir University of Technology; Tehran, Iran Email: momtazi@aut.ac.ir

Iranian Journal of Information Processing and Management

Received: 15, Jun. 2019 **Accepted: 15, Sep. 2020**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 36 | No. 3 | pp. 767-790
Spring 2021

Abstract: The progress of communications over internet media such as social media and messengers has led to the production of large amount of textual data. This kind of information contains a lot of valuable knowledge and can be used to improve the performance of other natural language processing (NLP) tasks. There are several ways to use such information, of which one is text summarization.

Summarizing textual information can extract the main content of text within a short time. In this paper, we propose an approach for extractive summarization on Persian texts by using sentences embedding and a sparse coding framework.

Most previous works focuses on text's sentences individually which may not consider the hidden structure patterns between them. In this paper, our proposed approach can consider the relations between the text's sentences by using three main criteria, namely coverage, diversity and sparsity, when selecting the summary sentences. By considering these criteria, we select sentences that can reconstruct the whole text with least reconstruction error.

The proposed approach is evaluated on Persian dataset Pasokh and achieved 10.02% and 8.65% improvement compared to the state-of-the-art methods in rouge-1 and rouge-2 f-scores, respectively. We show that considering semantic relations among the text's sentences can lead us to better sentence summarization.

Keywords: Text Summarization, Natural Language Processing, Sentence Embedding, Sparse Coding

* Corresponding Author

خلاصه‌سازی متون فارسی با استفاده از رویکرد کدگذاری تُنُک و بازنمایی عصبی جملات

رامین فتوره‌چی
کارشناسی ارشد هوش مصنوعی؛ دانشکدهٔ مهندسی کامپیوتر؛ دانشگاه صنعتی امیرکبیر؛ تهران، ایران؛
ramin.fatourechi@aut.ac.ir

سعیده ممتازی
دکتری هوش مصنوعی؛ استادیار؛ دانشکدهٔ مهندسی کامپیوتر؛ دانشگاه صنعتی امیرکبیر؛ تهران، ایران؛
پدیدآور رابط momtazi@aut.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۸/۰۳/۲۵ | پذیرش: ۱۳۹۹/۰۶/۲۵ | مقاله برای اصلاح به مدت ۱۰ ماه نزد پدیدآوران بوده است.

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۶ | شمارهٔ ۳ | ص‌ص ۷۶۷-۷۹۰
بهار ۱۴۰۰

چکیده: امروزه، گستردگی و تنوع اطلاعات متنی باعث پیچیدگی فرایند یافتن دانش و الگوهای مورد نظر از میان آن‌ها شده است. یکی از گام‌های مؤثر برای کاهش این مشکل، خلاصه‌سازی است. در چند دههٔ گذشته مسئلهٔ خلاصه‌سازی با توجه به نمونه‌های گوناگون از جهات و ابعاد مختلف بررسی شده است.

خلاصه‌سازی فرایندی هوشمند است که انجام آن حتی برای انسان‌ها ساده نیست و هر فردی با توجه به دیدگاهش می‌تواند نتیجهٔ متفاوتی ارائه دهد. یک خلاصهٔ مناسب باید دارای سه ویژگی پوشش، تُنُک‌بودن و تنوع باشد. بدین‌منظور در این پژوهش برای در نظر گرفتن این ویژگی‌ها یک روش بر مبنای کدگذاری تُنُک ارائه می‌شود. با به‌کارگیری این روش جملاتی به‌عنوان خلاصهٔ نهایی انتخاب می‌شوند که حداقل خطا را در بازسازی جملات متن ورودی داشته باشند. سپس، با استفاده از روش‌های عصبی در بازنمایی معنایی کلمات و همچنین متون به بهبود روش پیشنهادی پرداخته می‌شود. برای ارزیابی روش پیشنهادی از مجموعهٔ دادگان پاسخ استفاده شده و نشان داده می‌شود که روش پیشنهادی عملکرد بهتری نسبت به سایر پژوهش‌های انجام‌شده بر روی این دادگان در زبان فارسی دارد. مدل پیشنهادی توانسته است به میزان ۱۰/۰۲ درصد و ۸/۶۵ درصد و به‌ترتیب در معیار F روژ-۱ و روژ-۲ بهبود حاصل نماید.

کلیدواژه‌ها: خلاصه‌سازی، پردازش زبان طبیعی، کدگذاری تُنُک، بازنمایی جملات

۱. مقدمه

در چندین سال اخیر، تولید محتوا با پیشرفت تکنولوژی سرعت چشم‌گیری پیدا کرده است. مهم‌ترین دلیل برای آن تأثیر پیشرفت تکنولوژی بر نحوهٔ عملکرد و سرعت منابع تولید محتواست. منابع تولید محتوا می‌تواند رسانه‌های مختلف باشند. امروزه با توسعهٔ به‌کارگیری شبکهٔ جهانی وب، شبکه‌های اجتماعی، وبلاگ‌ها و همچنین رسانه‌های خبری برخط نیز به منابع مهم برای تولید محتوا توسط کاربران عادی تبدیل شده‌اند.

حجم بالای اطلاعات باعث می‌شود که هر کاربری که به‌دنبال جست‌وجوی اطلاعات است، با نتایج مختلف و متنوعی روبه‌رو شود که انتخاب اطلاعات مفید از میان آن‌ها را دشوار می‌کند. به‌عبارتی، این فراوانی اطلاعات خود می‌تواند باعث سرریز اطلاعات بوده و در نتیجه، موجب شود سرعتِ یافتن اطلاعات در حین فرایند جست‌وجو و تحلیل کاهش یابد. برای سرعت‌بخشیدن به فرایند کسب اطلاعات مفید و مختصر از منابع مختلف، طراحی یک سیستم خلاصه‌ساز از اجزای ضروری بسیاری از سامانه‌ها محسوب می‌شود. در سیستم‌های خلاصه‌ساز، هدف، استخراج اطلاعات مفید و کارا از میان مجموعهٔ بزرگی از اسناد با موضوع مشترک است که دارای کمترین میزان افزونگی اطلاعات باشد. فرایند تولید خودکار خلاصه به تقریب از نیم قرن پیش شروع شده و یکی از حوزه‌های فعال پژوهشی و عملیاتی در حوزهٔ هوش مصنوعی و پردازش زبان طبیعی است. این امر در سال‌های اخیر با توجه به رشد تکنولوژی‌های جدید و دلایلی که گفته شد، اهمیت فراوانی پیدا کرده است.

امروزه، خلاصه‌سازی متن از مهم‌ترین حوزه‌های موجود در پژوهش‌های پردازش زبان طبیعی است. با توسعهٔ روش‌های خلاصه‌سازی خودکار متن می‌توان هدف اصلی و کلیت متون و اسناد را در کمترین زمان ممکن دریافت. روش‌های مختلف خلاصه‌سازی می‌توانند کاربردهای مختلفی داشته باشد و با توجه به این کاربردها می‌توان روش خلاصه‌سازی مورد نظر را انتخاب نمود. از مهم‌ترین کاربردهای خلاصه‌سازی می‌توان به این موارد اشاره نمود: ۱) خلاصه‌سازی صفحات وب برای نمایش در موتورهای جست‌وجو، ۲) خلاصه‌سازی اسناد خبری با استفاده از روش‌های خلاصه‌سازی چندسندی متون برای یافتن اخبار مهم‌تر حول یک موضوع مشخص توسط خبرگزاری‌ها، و ۳) خلاصه‌سازی اسناد و نامه‌ها در ادارات (Gholamrezazadeh, AminiSalehi & Gholamzadeh 2009).

خلاصه‌سازی خودکار روشی است که طی آن محتوای مهم یک متن به‌طور خودکار توسط کامپیوتر تشخیص داده می‌شود و با خلاصه‌سازی متن که توسط انسان‌ها صورت می‌گیرد و در آن مفاهیم عمیق‌تری در نظر گرفته می‌شود، متفاوت است (Kiyoumarsi & Rahimi Esfahani 2011). سیستم‌های خلاصه‌سازی بسته به کاربردشان می‌توانند انواع مختلفی داشته باشند. این دسته‌بندی بر اساس نوع ورودی، نوع خروجی، هدف و روشی که مورد استفاده قرار می‌گیرد، صورت می‌گیرد. شکل ۱، دسته‌بندی انواع مختلف سیستم‌های خلاصه‌ساز را نمایش می‌دهد (همان).

خلاصه‌سازی بر اساس نوع ورودی، به دو دسته تک‌سندی[1] و چندسندی[2] تقسیم می‌گردد. در خلاصه‌سازی تک‌سندی فقط خلاصه‌ای مطابق با یک سند ورودی تولید می‌شود، اما در خلاصه‌سازی چندسندی ورودی می‌تواند شامل چندین سند جدا از هم پیرامون یک موضوع مشخص باشد (Gholamrezazadeh, AminiSalehi & Gholamzadeh 2009).

این سیستم‌ها را می‌توان از لحاظ نوع خروجی نیز دسته‌بندی نمود. خروجی که توسط این سیستم‌ها تولید می‌شود، می‌تواند عیناً از میان جملات ورودی انتخاب گردد. به این روش، خلاصه‌سازی استخراجی[3] گفته می‌شود. اما اگر خلاصهٔ تولیدشده مانند روش قبل عیناً جملات ورودی را در خلاصهٔ نهایی استفاده نکند و خروجی ترکیبی از بخش‌های مختلف جملات ورودی سیستم باشد، به آن خلاصه‌سازی انتزاعی[4] گفته می‌شود. اگرچه خلاصه‌سازی انتزاعی شباهت بیشتری به عملکرد طبیعی انسان دارد، اما با توجه به پایین‌بودن دقت آن در مقایسه با روش‌های استخراجی کمتر مورد توجه سیستم‌های پردازش متن قرار گرفته است (همان).

سیستم‌های خلاصه‌سازی را می‌توان بر اساس عملکردشان به دو دستهٔ عمومی و یا مبتنی بر یک موضوع خاص و یا یک پرسش مشخص دسته‌بندی کرد. در خلاصه‌سازی عمومی صرفاً متن ورودی بدون در نظر گرفتن هیچ پرس‌وجوی خاصی بررسی و خلاصه‌سازی می‌شود و چکیده‌ای از آن به‌عنوان خروجی ارائه می‌گردد. در این روش هیچ فرض ابتدایی نسبت به محتوای متن وجود ندارد و نسبت به تمام قسمت‌های مختلف متن ورودی به‌صورت یکسان رفتار می‌شود. اما می‌توان این عمل را نیز بر اساس درخواست و پرسش کاربر یا مخاطب انجام داد. در این روش به آن قسمت‌هایی

1. single-document 2. multi-document 3. extractive
4. abstractive

از متن ورودی که برای مخاطب اهمیت بیشتری دارد، توجه بیشتری خواهد شد. در برخی کاربردها متن ورودی فقط بر روی یک حوزهٔ خاص است؛ به‌طور مثال، متون اقتصادی، اجتماعی و یا ورزشی که برای خلاصه‌سازی آن‌ها نیاز به یک دانش نسبی به آن موضوع مشخص است (همان).

از نظر روشی که برای خلاصه‌سازی متون انتخاب می‌شود، می‌توان دو دستهٔ کلیِ با نظارت و بدون نظارت را در نظر گرفت. مهم‌ترین قدم در خلاصه‌سازی استخراجی امتیازدهی جملات و رتبه‌بندی آن‌ها بر اساس این امتیازات است. با پیشرفت روش‌های پردازش زبان طبیعی و یادگیری ماشین، روش‌های متعددی برای رتبه‌بندی جملات ارائه گردیده است. این روش‌ها بسته به نیازشان به داده‌های متنی برچسب‌گذاری‌شده، به دو دستهٔ با نظارت و بدون نظارت تقسیم می‌شوند. بنابراین، از نظر روشی که برای خلاصه‌سازی متون انتخاب می‌شود، می‌توان دو دستهٔ کلیِ با نظارت و بدون نظارت را در نظر گرفت. در خلاصه‌سازیِ با نظارت با یک مسئلهٔ طبقه‌بندی باینری روبه‌رو هستیم. این است که برای حل این مسئله نیاز به داده‌های متنی برچسب‌گذاری‌شده در مقیاس زیادی است. سپس، با آموزش مدلی بر روی این داده‌ها می‌توان ویژگی‌های متنی مورد نظر را برای انتخاب جملات خلاصه استفاده نمود. از آنجا که برچسب‌گذاری داده‌های خلاصه‌سازی نیاز به زمان و هزینهٔ بالایی دارد، از یادگیری بدون نظارت که به داده‌های برچسب‌گذاری‌شده نیازی ندارد، استقبال بیشتری می‌شود (Mao et al. 2019).

شکل ۱. انواع خلاصه‌سازی

در پژوهش حاضر خلاصه‌سازی تک‌سندی، استخراجی، به‌صورت عمومی و بدون ناظر مد نظر قرار گرفته تا با پیکره‌های موجود جهت ارزیابی سیستم‌های خلاصه‌سازی فارسی مطابقت داشته باشد.

پژوهش‌های مختلفی در سالیان اخیر بر روی خلاصه‌سازی متون برای زبان‌های مختلف صورت گرفته است. در زبان فارسی اغلب رویکردها با استفاده از ویژگی‌های آماری متون انجام گرفته و کمتر به ویژگی‌های معنایی متون توجه شده است. همچنین، طیف رویکردهایی که مورد استفاده قرار گرفته‌اند، زیاد گسترده نیست. با توجه به پیشینهٔ خلاصه‌سازی فارسی که در بخش ۲ به آن پرداخته می‌شود، در پژوهش حاضر، هدف، ارائهٔ روشی برای خلاصه‌سازی است که بر خلاف روش‌های آماری با الگوریتم‌های پیشرفته به حل مسئلهٔ خلاصه‌سازی پرداخته و به‌جای حجم بالایی از مهندسی ویژگی، از روش‌های سطح بالاتر مبتنی بر معنا استفاده شود. در همین راستا با الهام از پژوهش انجام‌شده توسط Liu, Yu & Deng (2015) روش کدگذاری تُنُک را برای این منظور مورد استفاده قرار می‌دهیم و به ارائهٔ مدل‌های مختلف برای بازنمایی متون جهت خلاصه‌سازی خواهیم پرداخت. بر این اساس، با به‌کارگیری روش‌های مبتنی بر شبکهٔ عصبی ویژگی‌های معنایی جملات استخراج می‌شود تا موجب بهبود روش‌های قبلی گردد.

۲. پیشینهٔ پژوهش

یکی از اولین پژوهش‌های خلاصه‌سازی انجام‌شده بر روی زبان فارسی، فارسی‌سام[1] است. این سیستم از ۳ فاز تشکیل شده است. در فاز اول، متن مورد نظر به سیستم داده می‌شود. سپس، با اعمال پیش‌پردازش بر روی آن‌ها جملات مشخص شده و کلمات توقف و اضافی از آن‌ها حذف می‌گردد. در ادامه، با استفاده از روش‌ها و ویژگی‌های زبانی، آماری و ابتکاری به جملات امتیاز تخصیص داده می‌شود. در فاز بعدی، امتیاز جملات نیز به‌عنوان معیاری دیگر به امتیاز جملهٔ مرتبط با آن اضافه می‌گردد و در فاز آخر، جملات با توجه به امتیازهایی که گرفته‌اند، رتبه‌بندی می‌شوند و خلاصهٔ مورد نظر با توجه به طول خلاصهٔ نهایی تولید می‌گردد (Hassel & Mazdak 2004).

در روش ارائه‌شده توسط Shamsfard & Karimi & Shamsfard (2006) با استفاده از روش‌های زنجیرهٔ واژگان و ویژگی‌های گراف و همچنین، با در نظر گرفتن ویژگی‌هایی

1. Farsi-Sum

چون شباهت بین جملات، شباهت به‌عنوان و ویژگی‌های آماری لغات، امتیاز جملات برای قرارگرفتن در خلاصهٔ نهایی محاسبه می‌شود. این روش برای خلاصه‌سازی تک‌سندهٔ متون استفاده شده است. ایدهٔ استفاده از گراف در خلاصه‌سازی متون در پژوهش دیگری توسط «حسین‌خواه، احمدی و محبی» ارائه گردیده است. در این پژوهش جهت ساخت بردار جملات و ایجاد گراف پردازش‌های زبان طبیعی از جمله تشخیص مقولهٔ دستوری کلمات و ریشه‌یابی کلمات انجام پذیرفته و سپس، گراف جملات بر مبنای شباهت بین جملات ایجاد شده است (Hosseinikhah, Ahmadi & Mohebi 2018).

در پژوهشی دیگر که توسط «حسین‌خواه، احمدی و محبی» ارائه گردید، با استفاده از ویژگی‌های متنی و رویکرد یادگیری ماشین خلاصهٔ نهایی انتخاب می‌شود. در این روش ابتدا برای بازنمایی جملات از ویژگی‌هایی همچون میانگین بسامد واژه- معکوس بسامد جمله[1]، طول جمله، جایگاه جمله، شباهت نسبت به‌عنوان، شباهت به کلمات کلیدی، انسجام بین جملات و تعداد رخداد واحدهای متنی خاص استفاده می‌گردد. سپس، با استفاده از منطق فازی و ویژگی‌های اشاره‌شده به ازای هر جمله، امتیاز بین صفر تا یک در نظر گرفته می‌شود. این امتیاز درجهٔ اهمیت آن جمله را نسبت به خلاصهٔ نهایی نشان می‌دهد (همان ۲۰۱۱).

همچنین، در مقالهٔ «توفیقی، راج و حاج سید جوادی» هر جمله با استفاده از ۶ ویژگی تعداد رخداد کلمات، وجود کلمات کلیدی، وجود کلمه‌ای از عنوان در جمله، جایگاه جمله و طول جمله نمایش داده می‌شود. سپس، با استفاده از روش فرایند سلسله‌مراتبی تحلیلی[2] که یک رویکرد برای تصمیم‌گیری با در نظر گرفتن چند معیار مختلف است، جملات مهم‌تر برای خلاصهٔ نهایی انتخاب می‌شوند. این روش با در نظر گرفتن میزان تأثیر هر معیار میزان اهمیت هر جمله را مشخص می‌کند (Tofighy, Raj & HajSeyyedJavadi 2013).

در پژوهشی که توسط «سلطانی، نصیری و عسگریان» ارائه شد نیز امتیازدهی برای انتخاب جملات خلاصه مورد استفاده قرار گرفته است. در این پژوهش امتیازدهی بر اساس ویژگی‌های آماری مانند طول جمله، موقعیت جمله، شباهت با زمینهٔ متن و ویژگی‌های زبانی مانند عبارات و اسامی خاص انجام پذیرفته است (Soltani, Nasiri & Asgarian 2018). در مقاله‌ای دیگر با تبدیل هر سند ورودی به ماتریس (عبارت-جمله) و با استفاده از روش تجزیهٔ

1. Term Frequency-Inverse Sentence Frequency (TF-ISF)
2. analytic hierarchy process (AHP)

مقدارهـای منفـرد[1] سـه ماتریـس V، Σ، و U به‌دسـت می‌آیـد. بـا به‌کارگیـری ماتریـس Σ تعـداد مفاهیـم مهـم سـند ورودی اسـتخراج می‌شـود و سـپس، بـا اسـتفاده از ماتریـس V میـزان اهمیـت جمـلات سـند ورودی اسـتخراج می‌شـود (Honarpisheh, Ghassem-Sani & Mirroshandel). در پژوهشـی دیگـر، درجـهٔ اهمیـت کلمـات اسـناد ورودی بـا اسـتفاده از روش‌هـای زنجیـرهٔ واژگان تخمیـن زده می‌شـود. سـپس، گرافـی از ایـن کلمـات تشـکیل می‌گـردد کـه کلمـات به‌عنـوان گره‌هـای گـراف در نظـر گرفتـه شـده و یال‌هـای گـراف بـا توجـه بـه میـزان اهمیتـی کـه بیـن کلمـات وجـود دارد، در نظـر گرفتـه می‌شـوند. سـپس، امتیـاز جمـلات سـند ورودی بـا اسـتفاده از همیـن گـراف و وزن یـال بیـن کلمـات گـراف (گره‌هـا) تخمیـن زده می‌شـود و آنـگاه جمـلات مهم‌تـر بـرای خلاصـهٔ نهایـی یافـت می‌شـوند (Shakeri et al. 2012). همچنیـن، در روشـی مشـابه بـا اسـتفاده از گـراف، جمـلات به‌عنـوان گره‌هـای گـراف در نظـر گرفتـه می‌شـود. در ایـن گـراف بـرای هـر گـره و هـر یـال یـک میـزان اهمیـت در نظـر گرفتـه می‌شـود. میـزان اهمیـت هـر یـال بـا توجـه بـه میـزان ارتبـاط دو جملـهٔ متناظـر بـا آن یـال تخمیـن زده می‌شـود. همچنیـن، میـزان اهمیـت هـر گـره نیـز بـا توجـه بـه معیارهایـی ماننـد تعـداد یال‌هـای ورودی بـه گـره، تعـداد رخـداد کلمـات و تعـداد کلمـات کلیـدی در جملـهٔ متناظـر به‌دسـت می‌آیـد. در نهایـت، بـا اسـتفاده از وزن‌هـای به‌دسـت‌آمده جمـلات مهم‌تـر مشـخص می‌شـوند (همـان).

در روشـی دیگـر تحـت عنـوان «پارسومیسـت»[2] بـا به‌کارگیـری روش‌هـای ترکیبـی آمـاری، معنایـی و ابتـکاری خلاصـهٔ نهایـی یـک سـند و یـا چنـد سـند را تولیـد می‌کننـد. ایـن روش شـامل سـه بخـش پیش‌پـردازش، تحلیـل و انتخـاب اسـت. بـا دریافـت یـک سـند ورودی ابتـدا بـا اسـتفاده از منابـع خارجـی بـه پیش‌پـردازش آن می‌پردازنـد و سـپس، بـه هـر جملـهٔ آن یـک امتیـاز اختصـاص می‌دهنـد و در آخـر، جمـلات مهم‌تـر و غیـر تکـراری را به‌عنـوان خلاصـهٔ نهایـی در نظـر می‌گیرنـد (Shamsfard, Akhavan & Erfani Jourabchi 2009). در مقالـهٔ «خادمـی» و همـکاران روشـی بـرای خلاصه‌سـازی اسـتخراجی متـون بـا در نظـر گرفتـن مفاهیـم موجـود در جمـلات پیشـنهاد گردیـده اسـت. در ایـن روش مفاهیـم موجـود در هـر سـند بـا اسـتفاده از روش‌هـای بازنمایـی کلمـات خوشـه‌بندی می‌شـوند. سـپس، ایـن خوشـه‌های مفاهیـم امتیازدهـی می‌شـوند. میـزان اهمیـت هـر جملـه بـا توجـه بـه میـزان اهمیـت مفاهیـم استخراج‌شـده و موجـود در آن به‌دسـت می‌آیـد (Khademi, Fakhredanesh & Hoseini 2017). در روشـی دیگـر تحـت عنـوان «تب–سـام»[3] بـا اسـتفاده از روشـی مشـابه، پـس از پیش‌پـردازش متـن ورودی و بـا اسـتفاده

1. singular value decomposition (SVD) 2. Parsumist 3. TabSum

از ویژگی‌های آماری متن ورودی و منابعی مانند زنجیرهٔ واژگان به خلاصه‌سازی تک‌سندی متن پرداخته شده است (Masoumi, Feizi-Derakhshi & Tabatabaei 2014). در پژوهش «توفیقی» و همکاران با تجزیهٔ متن ورودی به یک ساختار سلسله‌مراتبی و ساخت درخت فرکتال[1] مجدداً به استخراج ویژگی‌های آماری در سطح کلمه و جمله پرداخته شده و سپس، به محاسبهٔ امتیاز اجزای سند ورودی با استفاده از این ویژگی‌ها اقدام می‌شود (Tofighy et al. 2011). در پژوهش «روحانیان» خلاصه‌سازی چندسندهٔ متون مورد توجه قرار گرفته است. این بار در این روش به بازنمایی پاراگراف‌ها در فضای برداری پرداخته می‌شود. سپس، با استفاده از الگوریتم خوشه‌بندی به جداسازی این پاراگراف‌های اسناد ورودی پرداخته می‌شود و قسمت‌های مهم‌تر اسناد ورودی مشخص می‌گردند (Rohanian 2017).

یک خلاصهٔ مطلوب خلاصه‌ای است که با یک بار خواندن آن بتوان کل مفهوم و پیام اسناد مورد نظر را متوجه شد. بیشتر مدل‌های آماری نمی‌توانند تمامی جنبه‌های موجود در یک متن را در نظر بگیرند، زیرا هر جمله را نسبت به خود آن در نظر گرفته و رابطهٔ آن با سایر جملات را از نظر معنایی در نظر نمی‌گیرند. در مقالهٔ حاضر سعی شده است جملات متن با ارائهٔ یک رویکرد مناسب با استفاده از بازنمایی عصبی جملات و کدگذاری تُنُک به بهترین شکل ممکن بازنمایی شوند و از میان آن‌ها جملاتی که جملات دیگر متن ورودی را به بهترین شکل ممکن بازسازی می‌کنند انتخاب می‌شوند.

دو ویژگی اصلی روش پیشنهادی استفاده از روشی است که ۱) ابعاد مختلف پوشش، تُنُک بودن و تنوع را در خلاصه‌سازی لحاظ کند، و (۲) به‌جای مهندسی ویژگی و وابستگی به پردازش‌های مختلف متنی، با یک رویکرد یکپارچه به بازنمایی معنایی متون بپردازد.

۳. مدل خلاصه‌سازی مبتنی بر کدگذاری تُنُک

همان‌طور که اشاره شد، بسیاری از روش‌های موجود به‌منظور خلاصه‌سازی استخراجی از مدل‌های رتبه‌بندی و امتیازدهی بر اساس ویژگی‌های آماری متن مورد نظر برای انتخاب جملات برتر استفاده می‌کنند. اما این روش‌ها ممکن است جملاتی را انتخاب کنند که معنا و مفهوم تکراری داشته باشند. با توجه به این چالش، در پژوهش حاضر روشی ارائه می‌شود که تا حد ممکن این مشکل را برطرف نماید. در این روش ابتدا جملات متن ورودی پس از پیش‌پردازش با استفاده از روش‌های بازنمایی جملات به

1. fractal

فضای برداری نگاشت می‌شوند. سپس، با استفاده از الگوریتم کدگذاری تُنُک، محتوای جملات و ارتباط آن‌ها نسبت به هم بررسی می‌شود و خلاصهٔ نهایی تولید می‌گردد. در این جا روش پیشنهادی دارای سه قسمت اصلی پیش‌پردازش متن ورودی، بازنمایی جملات و تولید خلاصهٔ نهایی است که این سه مورد در ادامه توضیح داده خواهد شد. بنا بر دانش نویسندگان پژوهش حاضر اولین کار برای خلاصه‌سازی متون فارسی با رویکرد کدگذاری تُنُک است.

۳-۱. پیش‌پردازش متن ورودی

در این بخش، متن مورد نظر برای خلاصه‌سازی از کلمات اضافی مانند ایست‌واژه‌ها[1] و علائم متن پاک‌سازی می‌شود. سپس، جملات متن ورودی تشخیص داده می‌شوند و از هم جدا می‌گردند. در این پژوهش چون از روش خلاصه‌سازی استخراجی استفاده شده، تشخیص صحیح جملات متن ورودی از یکدیگر امری مهم است. در نتیجه، جملات برای نگاشت به فضای برداری برای ادامهٔ عملیات آماده می‌شوند.

۳-۲. بازنمایی جملات

برای بسیاری از روش‌های پردازش متن، نیاز به نمایش عددی کلمات و متون است تا بتوان از انواع رویکردهای موجود در حوزهٔ یادگیری ماشین مانند الگوریتم‌های دسته‌بندی روی لغات و اسناد استفاده نمود. یکی از رهیافت‌هایی که در این حوزه بسیار رایج شده، نمایش برداری کلمات و جملات است. افزایش محبوبیت و کاربرد شبکه‌های عصبی در چند سال اخیر موجب پیشرفتی چشمگیر در حوزهٔ پردازش متن گردیده است. از جملهٔ مهم‌ترین این کاربردها، در زمینهٔ توسعهٔ مدل‌های زبانی است که به‌تبع آن روش‌های بازنمایی متون مانند «ورد۲وک»[2] (Mikolov et al. 2013) معرفی شده است. همان‌طور که می‌دانید، بازنمایی واحدهای زبانی (جمله، کلمه، سند و ...) یکی از مهم‌ترین نیازهای حوزهٔ پردازش زبان طبیعی است. در این پژوهش برای بازنمایی جملات دو روش به‌کار رفته است:

۱. روش بسامد واژه-معکوس بسامد سند[3]؛
۲. روش میانگین بردارهای «ورد۲وک» کلمات جملهٔ مورد نظر.

1. stop words 2. Word2Vec 3. term frequency-inverse document frequency (TF-IDF)

بازنمایی جملات با استفاده از روش «بسامد واژه-معکوس بسامد سند»، یکی از روش‌های پایه در پردازش متون است. در این روش هر واحد از بردار بازنمایی مورد نظر به یک کلمهٔ مشخص تخصیص داده می‌شود و معرف میزان تکرار کلمه در سند با وزنی معادل معکوس تعداد اسناد حاوی آن کلمه است. این روش، نوعی بازنمایی محلی است که به‌روش کدگذاری تک‌روشن[1] معروف است.

در بازنمایی کلمات با استفاده از «ورد۲وک» هر بعد از بردارهای بازنمایی نشان‌دهندهٔ یک ویژگی استخراج‌شده از کلمهٔ مورد نظر است. برای داشتن این‌گونه مدلی برای بازنمایی کلمات، نیاز به طی فرایندی شبیه به ساخت مدل زبانی بر روی دادهٔ متنی مورد نظر است. به‌عبارت دیگر، به مدلی احتیاج هست که بتواند با داشتن یک واژه کلمات اطراف آن را پیش‌بینی کند.

بازنمایی کلمات در روش «ورد۲وک» به دو صورت کیسهٔ کلمات پیوسته[2] و پرش-نگاشت[3] امکان‌پذیر است. در روش اول، با داشتن کلمات، بافت کلمهٔ هدف پیش‌بینی می‌شود و در روش دوم، با داشتن کلمهٔ هدف، کلمات بافت پیش‌بینی می‌شود. در پژوهش حاضر روش پرش-نگاشت به‌کار رفته است.

برای بازنمایی جملات با استفاده از این مدل، ابتدا می‌بایست بردارهای کلمات جملهٔ مورد نظر را استخراج نمود. سپس، برای تعیین بردار نهایی جملهٔ مورد نظر، می‌توان میانگین بردارهای استخراج‌شده را به‌عنوان بازنمایی مورد نظر مشخص نمود.

در نتیجه، جملات متن ورودی که از مرحلهٔ قبل آماده‌سازی شده بودند، با استفاده از روش بازنمایی گفته‌شده به فضای برداری انتقال داده می‌شوند و یک ماتریس از بردار جملات برای مرحلهٔ بعدی روش خلاصه‌سازی مورد نظر آماده می‌گردد. هر سطر این ماتریس، بردار هر یک از جملات متن ورودی است.

۳-۳. الگوریتم خلاصه‌سازی

یک خلاصهٔ مطلوب خلاصه‌ای است که با یک بار خواندن آن بتوان کل مفهوم و پیام متن مورد نظر را متوجه شد. مدل‌های رتبه‌بندی اغلب نمی‌توانند تمامی جنبه‌های موجود در یک متن را در نظر بگیرند، زیرا امتیازی که به هر جمله می‌دهند، تنها نسبت به همان جمله است و رابطهٔ میان جمله‌ها در نظر گرفته نمی‌شود. در واقع، جملاتی که

1. one-hot
2. continuous bag of words (CBOW)
3. skip-gram

به‌عنوان خلاصهٔ نهایی انتخاب می‌شوند باید این قابلیت را داشته باشند که جملات دیگر متن ورودی از روی آن‌ها بازسازی شود. در نتیجه، یک خلاصهٔ کامل و جامع می‌بایست سه ویژگی زیر را داشته باشد (Kiyoumarsi & Rahimi Esfahani 2011):

- ◇ پوشش[1]؛
- ◇ تُنُک بودن[2]؛
- ◇ تنوع[3].

منظور از پوشش این است که جملات انتخاب‌شده به‌عنوان خلاصهٔ نهایی، شامل تمام جنبه‌های مختلف متن ورودی باشند؛ یعنی فقط شامل جملات مهم متن ورودی نباشند و تمام جنبه‌های مجموعهٔ جملات متن ورودی مورد نظر در آن پوشش داده شود. بدین جهت در این جا از یک ترکیب خطی غیر منفی برای نمایش رابطه میان مجموعهٔ جملات متن ورودی و جملات نهایی خلاصهٔ تولیدشده استفاده شده است. از طریق این رابطه می‌توان مجموعهٔ جملات متن ورودی را با استفاده از جملات خلاصهٔ نهایی بازسازی نمود.

تُنُک‌بودن به این معناست که هر یک از جملات متن ورودی توسط تعداد کمی از جملات خلاصهٔ نهایی نمایش داده می‌شوند. مجموعهٔ جملات متن ورودی معمولاً شامل یک عنوان اصلی و چندین عنوان فرعی دیگر هستند که جملات خلاصه می‌بایست به این عناوین تعلق داشته باشند. در واقع، هر جمله از متن ورودی تا حدی محتوای دیگر جملات را در خود نمایش می‌دهد. هر جمله‌ای که محتوای تعداد جملات بیشتری را بتواند به‌صورت بهتری در خود داشته باشد، ارزش انتخاب به‌عنوان خلاصهٔ نهایی را دارد.

برای ایجاد تنوع میان جملات خلاصه باید از انتخاب جملات تکراری به‌عنوان خلاصهٔ نهایی جلوگیری شود. یک خلاصه فقط نباید موضوع اصلی مجموعهٔ اسناد ورودی را در خود داشته باشد. این کار باعث می‌شود که جملاتی که انتخاب می‌شوند، مفهوم تکراری و شبیه به هم داشته باشند. برای جلوگیری از انتخاب جملات تکراری پیشنهاد می‌شود که علاوه بر در نظر گرفتن موضوع اصلی متن ورودی، عناوین فرعی آن نیز برای انتخاب جملات خلاصه در نظر گرفته شود. به این منظور، در این جا همبستگی زوج جملات خلاصه که کمترین تفاوت را با هم دارند، به‌عنوان معیاری از تنوع در نظر گرفته می‌شود؛ یعنی سعی می‌شود فاصله میان زوج جملات خلاصه با هم تا حد ممکن

1. coverage 2. sparsity 3. diversity

زیاد باشد تا از انتخاب جملات تکراری و شبیه به هم جلوگیری گردد. همچنین، اگر از ویژگی تنوع در ساختِ خلاصه استفاده شود، می‌توان کل دیدگاه متن ورودی را در خلاصهٔ نهایی بیان نمود.

بر اساس این سه ویژگی، با الهام از پژوهش انجام‌شده توسط Liu, Yu & Deng (2015) روش کدگذاری تُنُکی که در ادامه توضیح داده خواهد شد، برای خلاصه‌سازی متن مورد استفاده قرار می‌گیرد. این رویکرد دارای دو سطح است:

سطح اول: مجموعه خلاصه بازنمایی تُنُکی از متن ورودی است؛

سطح دوم: هر جمله از متن ورودی نیز به‌صورت تُنُک توسط جملات مجموعهٔ خلاصه بازسازی شده است. یعنی برای بازسازی هر جمله از یک‌سری از جملات مجموعهٔ خلاصه استفاده می‌شود.

هر جمله از جملات متن ورودی می‌تواند از طریق یک ترکیب خطی وزن‌دار غیرمنفی مجموعهٔ جملات خلاصه تخمین زده شود که به‌صورت رابطهٔ (۱) به‌دست می‌آید:

$$s_i \approx \sum_{j=1}^{k} a_{ji} s_j^* \qquad (۱)$$

$$s.t. \quad a_{ji} \geq 0$$

که در آن a_{ji} ضریب ترکیب خطی میان جملهٔ jام از مجموعهٔ جملات خلاصه به جملهٔ iام از مجموعهٔ جملات کاندید است و مقدار این ضریب همیشه بزرگ‌تر یا برابر با صفر است.

برای ارزیابی میزان پوشش خلاصه مقدار خطای بازسازی با استفاده از L2-norm به‌صورت زیر تعریف می‌شود:

$$re(s_i) = \left\| s_i - \sum_{j=1}^{k} a_{ji} s_j^* \right\|_2^2 \qquad (۲)$$

در نتیجه، تابع هزینه تا این جا برابر با خطای بازسازی مجموعهٔ جملات کاندید با استفاده از مجموعهٔ خلاصه است:

$$J = \min_{S^*,A} \sum_{i=1}^{n} \left\| s_i - \sum_{j=1}^{k} a_{ji} s_j^* \right\|_2^2 \qquad (۳)$$

$$s.t. \quad a_{ji} \geq 0$$

برای اطمینان از این‌که هر یک از جملات توسط تعدادی از جملات خلاصه بازسازی شده است، محدودیت تُنُک‌بودن بر روی ستون‌های ماتریس A (ماتریس ضرایب) اعمال شده است. این محدودیت با استفاده از L1-norm به‌صورت زیر به تابع هزینه رابطهٔ (۴) اضافه گردیده است و می‌توان مقدار تأثیر آن را با تغییر ضریب کنترل نمود.

$$J = \min_{S^*,A} \sum_{i=1}^{n} \left\| s_i - \sum_{j=1}^{k} a_{ji} s_j^* \right\|_2^2 + \lambda \sum_{i=1}^{n} \|a_{:i}\|_1 \qquad (۴)$$

$$s.t. \quad a_{ji} \geq 0, \ \lambda > 0$$

همچنین، برای تأثیر ویژگی متنوع‌بودن مجموعهٔ جملات خلاصه، به تابع هزینهٔ (۴) عبارت حداکثر امتیاز همبستگی اضافه گردیده است که حاصل آن رابطهٔ (۵) است.

$$J = \min_{S^*,A} \sum_{i=1}^{n} \left\| s_i - \sum_{j=1}^{k} a_{ji} s_j^* \right\|_2^2 + \lambda \sum_{i=1}^{n} \|a_{:i}\|_1 + \beta \max_{j \neq k} corr(s_j^*, s_k^*) \qquad (۵)$$

$$s.t. \quad a_{ji} \geq 0, \ \lambda > 0, \ \beta > 0$$

همچنین، مقدار تابع همبستگی فوق به‌صورت زیر تعریف می‌شود:

$$corr(s_i^*, s_j^*) = \frac{(s_i^* - \overline{s_i^*})(s_j^* - \overline{s_j^*})}{\sigma_i \sigma_j} \qquad (۶)$$

تابع هزینهٔ به‌دست‌آمده از روابط فوق یک مسئلهٔ NP-hard است که حل آن میسر نیست. بدین‌منظور، برای حل این تابع هزینه از الگوریتم تبرید شبیه‌سازی‌شده[1] برای یافتن بهترین جواب استفاده شده است. در الگوریتم ۱، جزئیات الگوریتم پیشنهادی شرح داده شده است (Liu, Yu & Deng 2015).

1. simulated annealing

الگوریتم اول

- ورودی‌ها: مجموعهٔ جملات ورودی S ، تعداد جملات خلاصهٔ نهایی k، ضریب تُنُک بودن $\boldsymbol{\lambda}$، ضریب همبستگی $\boldsymbol{\beta}$، $k=0$، تابع هزینه J ، خلاصهٔ ابتدایی به‌صورت تصادفی انتخاب می‌گردد. $\boldsymbol{S^*}$، دمایی که در آن الگوریتم تبرید شبیه‌سازی‌شده توقف می‌کند $\boldsymbol{T_{stop}}$
- خروجی: خلاصهٔ نهایی $\boldsymbol{S^*}$

مراحل الگوریتم:

- تا زمانی که $\boldsymbol{T_k > T_{stop}}$ باشد:
  - ماتریس A توسط تابع کدگذاری تُنُک تخمین زده خواهد شد. (تابع *Sparse-coding* در الگوریتم دوم توضیح داده خواهد شد)
  - اگر $\boldsymbol{J(S,A,T_k) < J_{opti}}$ باشد:
    - $\boldsymbol{J_{opti} \leftarrow J(S,A,T_k)}$ و $\boldsymbol{S^*_k \leftarrow S^*_{opti}}$ می‌شود.
  - در غیر این صورت:
    - $\boldsymbol{rej \leftarrow rej + 1}$
  - و اگر $\boldsymbol{ref < MaxConseRej}$ بود:
    - $\boldsymbol{S^*_{opti}}$ به‌عنوان خروجی الگوریتم برگشت داده می‌شود.
  - برای هر جمله $\boldsymbol{s^*}$ موجود در مجموعهٔ جملات خلاصه $\boldsymbol{S^*_k}$:
    - $\boldsymbol{tmp \leftarrow Update_{S^*}(s^*, T_k)}$
    - اگر به‌روزرسانی انجام‌شده مورد قبول واقع شد:
      - $\boldsymbol{S^*_{k+1} \leftarrow S^*_{k+1} \cup tmp}$
    - در غیر این صورت:
      - $\boldsymbol{S^*_{k+1} \leftarrow S^*_{k+1} \cup s^*}$
  - $\boldsymbol{T_{k+1} \leftarrow Update\_T(k)}$
  - $\boldsymbol{k \leftarrow k + 1}$
- پس از پایان الگوریتم، $\boldsymbol{S^*_{opti}}$ به‌عنوان خروجی الگوریتم برگشت داده می‌شود.

محدودهٔ جست‌وجوی تابع در تکرارهای ابتدایی به‌سرعت کاهش پیدا می‌کند، ولی هرچه دما پایین‌تر بیاید، نقاط محلی را بیشتر مورد جست‌وجو قرار می‌دهد.

برای انتخاب همسایهٔ هر جمله از مجموعه خلاصهٔ شبیه‌ترین جمله به آن به‌عنوان همسایه انتخاب می‌شود. در این حالت ممکن است جملهٔ انتخاب‌شده قبلاً به‌عنوان همسایهٔ یک جملهٔ دیگر انتخاب شده باشد. بدین‌منظور، برای این‌که جملهٔ تکراری وارد مجموعهٔ خلاصه نشود، دومین جملهٔ شبیه به‌عنوان همسایه انتخاب می‌شود. در کل، هر زمان جملهٔ تکراری انتخاب گردد، شبیه‌ترین جملهٔ بعدی به‌عنوان همسایهٔ جایگزین و مجموعهٔ خلاصه به‌روزرسانی می‌شود.

هر بار که مجموعهٔ خلاصه به‌روزرسانی شد، مقدار تابع هزینه با توجه به تغییر

داده‌شده، مجدد محاسبه می‌شود. اگر تغییر اعمال‌شده موجب کاهش تابع هزینه شده باشد، تغییر مورد قبول واقع می‌گردد. الگوریتم ۲، نحوهٔ محاسبهٔ ماتریس ضرایب A را نشان می‌دهد:

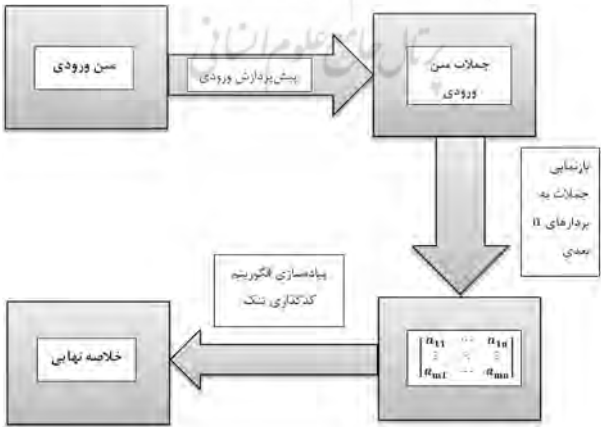
الگوریتم دوم: محاسبهٔ ماتریس ضرایب A

- ورودی‌ها: مجموعهٔ جملات کاندید S، جملات خلاصه S^*
- خروجی: ماتریس ضرایب A

مراحل الگوریتم:

- ماتریس A با استفاده از مقدار $\frac{1}{k}$ مقداردهی اولیه می‌شود.
- تا زمانی که $t < 100$ است:
  - $A^{t+1} = A^t .* (\overline{S^*}S)./(\overline{S^*}SA^t + \lambda 1)$
  - اگر $norm(A^{t+1} - A^t) < 0.01$ باشد:
    - مقدار A^t به‌عنوان جواب الگوریتم برگشت داده می‌شود.
  - در غیر این صورت:
    - $t \leftarrow t + 1$
- در پایان، حلقهٔ مقدار A^t به‌عنوان جواب الگوریتم برگشت داده می‌شود.

بدین‌صورت ماتریس بردارهای جملات که از قسمت قبل به‌دست آمده بود، توسط الگوریتم پیشنهادی مورد پردازش قرار می‌گیرد. در نهایت، بردارهای مهم‌تر ماتریس ورودی تشخیص داده می‌شوند. نتیجه این‌که، جملات متناظر با آن بردارها استخراج می‌شوند و خلاصهٔ نهایی تشکیل می‌گردد. شمای کلی ساختار مدل پیشنهادی در شکل ۲، نمایش داده شده است.



شکل ۲. ساختار روش پیشنهادی

٤. نتایج آزمایش‌ها

٤-١. دادگان

به‌منظور ارزیابی روش پیشنهادی از مجموعهٔ دادگان پاسخ استفاده شده است. مجموعهٔ دادگان پاسخ شامل تعداد زیادی از اسناد خبری به همراه خلاصهٔ متناظر با آن‌هاست که توسط انسان‌ها نوشته شده است. قبل از این مجموعهٔ دادگان، مجموعهٔ مناسبی برای توسعهٔ خلاصه‌سازی متون در زبان فارسی وجود نداشت. این مجموعهٔ دادگان مطابق استانداردهای جهانی به‌منظور ارزیابی خلاصه‌سازی تک‌سندی، چندسندی، استخراجی و انتزاعی تهیه شده است (Pour-masoomi 2014).

در این پژوهش از اسناد تک‌سندی و خلاصه‌های استخراجی آن به‌منظور ارزیابی روش ارائه‌شده استفاده خواهد شد. تعداد ۱۰۰ سند خبری با طول متفاوت از چندین مرکز خبری مختلف بدین منظور تهیه شده است. همچنین، متون این پیکره از شش حوزهٔ فرهنگی، اقتصادی، ورزشی، اجتماعی، سیاسی و علمی جمع‌آوری گردیده است. جدول ۱، خلاصهٔ اطلاعات این پیکره را نمایش داده است.

جدول ۱. جزئیات دادگان تک‌سندی پاسخ

مورد	تعداد
اسناد موجود در پیکره	۱۰۰
ژانر خبرها	۶
خبرگزاری‌ها	۷
خلاصه‌های استخراجی برای هر سند	۵
خلاصه‌های انتزاعی برای هر سند	۵

٤-٢. معیار ارزیابی

بین خلاصه‌هایی که توسط انسان‌ها تولید می‌شود، شباهت زیادی وجود ندارد و هر کس بنا بر دیدگاه و عقاید خود یک خلاصه را نگارش و ارزیابی می‌کند. سخت‌تر از ارزیابی قالب و فرم یک خلاصه، ارزیابی محتوای اصلی آن است. نکتهٔ مهم دیگر در ارزیابی خلاصه‌ها گسترده‌بودن پارامترها در ارزیابی است؛ یعنی به‌دست آوردن یک معیار واحد در این کار دشوار است. با در نظر گرفتن این چالش‌ها معیار ارزیابی

روژ[1] برای ارزیابی خلاصه‌سازی خودکار متون پیشنهاد گردیده است. این معیار خلاصهٔ تولیدشده را می‌تواند از جهات مختلف بررسی نماید. برای ارزیابی نتایج پژوهش حاضر از معیارهای روژ-۱[2] و روژ-۲[3] که جزء دستهٔ روژ-n[4] قرار دارند، استفاده خواهد شد.

در کل، روژ-n برابر است با مجموع تعداد مطابقت‌های nتایی‌ها بین خلاصهٔ ایده‌آل و خلاصهٔ تولیدشده تقسیم بر مجموع کل nتایی‌های موجود در خلاصهٔ ایدئال. در نتیجه، روژ-۱ همپوشانی‌های یکی‌ای‌ها[5] را بین خلاصهٔ تولیدشده و خلاصهٔ دستی مقایسه می‌کند و در روژ-۲ این مقایسه نسبت به دوتایی‌های مشترک است (Lin 2004).

این معیار همان‌طور که از مخرج آن معلوم است، مبتنی بر بازخوانی است. اما معیار «بلو»[6] که به‌عنوان یک معیار در ارزیابی ماشین ترجمه استفاده می‌شود، دقت‌محور است. این معیار بررسی می‌کند که یک مورد کاندید چقدر با مجموعهٔ ایده‌آل و مرجع مطابقت دارد.

در نظر داشته باشید که در روژ-n هرچه به تعداد خلاصه‌های ایده‌آل افزوده شود، مخرج آن بزرگ‌تر می‌شود. در واقع، هرچه به تعداد خلاصه‌های ایده‌آل موجود افزوده شود، می‌توان جنبه‌های مختلف خلاصهٔ تولیدشده را مورد بررسی قرار داد. در نتیجه، در روژ-n یک خلاصهٔ کاندید که تعداد nتایی‌های مشترک بیشتری با خلاصه‌های ایده‌آل داشته باشد، اهمیت و امتیاز بیشتری پیدا می‌کند.

٤-۳. تنظیم آزمایش‌ها

در انجام آزمایش‌ها جهت ساخت بردارهای «ورد۲وک» کلمات ۴ کلمهٔ قبل و ۴ کلمهٔ بعد از هر کلمه به‌عنوان بافت کلمه لحاظ گردیده و کلماتی که کمتر از ۵ بار تکرار شده‌اند، نادیده گرفته شده است.

در روش به‌کاررفته پارامترهای مورد نیاز، پارامترهای رابطهٔ (۵) است که شامل پارامترهای لامبدا و بتا برای تنظیم مقدار تابع هزینه است. برای اطمینان از این‌که هر یک از جملات توسط تعدادی از جملات خلاصه بازسازی شده است، محدودیت تُنُک‌بودن بر روی ستون‌های ماتریس A (ماتریس ضرایب) اعمال شده است و می‌توان مقدار تأثیر آن را با استفاده از ضریب لامبدا کنترل نمود. هرچه مقدار این ضریب بیشتر باشد، جملات

1. Rouge
2. Rouge-1
3. Rouge-2
4. Rouge-n
5. unigram
6. BLEU

ورودی تأثیر بیشتری در انتخاب جملات خلاصهٔ نهایی دارند، اما وقتی این تأثیر از حدی بگذرد، دیگر جملات مهم‌تر برای خلاصهٔ نهایی انتخاب نمی‌شوند. در پژوهش حاضر مقدار لامبدا برابر ۰/۵ لحاظ شده است. همچنین، برای تأثیر ویژگی متنوع‌بودن مجموع جملات خلاصه به تابع هزینه عبارت حداکثر میزان همبستگی نیز اضافه گردیده است که مقدار آن با استفاده از ضریب بتا کنترل می‌شود. این ضریب نیز هرچه بزرگ‌تر باشد، جملات خلاصهٔ نهایی ارتباط کمتری با یکدیگر خواهند داشت و اگر این مقدار نیز از حدی بگذرد، دیگر خلاصهٔ نهایی انسجام لازم از نظر مرتبط‌بودن جملات انتخاب‌شده را با یکدیگر ندارند. در پژوهش حاضر مقدار بتا برابر ۱ لحاظ شده است.

٤-٤. تحلیل نتایج

در این بخش عملکرد روش پیشنهادی این پژوهش برای خلاصه‌سازی متون فارسی بررسی خواهد شد. تاکنون بهترین نتایجی که بر روی دادگان فارسی پاسخ به‌دست آمده است، با استفاده از روش‌های بدون نظارت و استفاده از ویژگی‌های متنی بوده است.

در پژوهش انجام‌شده توسط «کهنسال، فیلی و فرضی» و همکاران رویکرد خوشه‌بندی و رویکرد مبتنی بر گراف TextRank مورد استفاده قرار گرفته است (۱۳۹۶). همچنین، برای بازنمایی جملات از مدل میانگین «ورد۲وک» و میانگین وزن‌دار «ورد۲وک»، میانگین تطابق، و نرمال تطابق استفاده شده است. در ادامه، جدولی از نتایج پژوهش «کهنسال، فیلی و فرضی» بر روی دادگان پاسخ نشان داده خواهد شد. همچنین، دو سامانهٔ «ایجاز»[1] Pour-masoomi et al. (2014) و «فارسی‌سام» Hassel & Mazdak (2004) که هر دو بر مبنای امتیازدهی عمل می‌کنند نیز بر روی دادگان پاسخ مورد ارزیابی قرار گرفته‌اند که نتایج آن‌ها در جدول ۲، آورده شده است. نتایج ارائه‌شده در این جدول به‌عنوان نتایج پایه از کارهای انجام‌شده در خلاصه‌سازی فارسی مد نظر قرار می‌گیرد و معیار مقایسه روش پیشنهادی در پژوهش حاضر است.

1. Ijaz

جدول ۲. نتایج کار های انجام‌شده در خلاصه‌سازی فارسی[1]

		روژ-۱			روژ-۲		
نوع مدل	نوع بازنمایی	دقت	فراخوانی	معیار F	دقت	فراخوانی	معیار F
TextRank	میانگین ورد۲وک	۳۴/۲۸	۵۷/۵۳	۴۲/۹۶	۲۳/۱۷	۳۹/۵۲	۲۹/۲۱
	میانگین وزن‌دار ورد۲وک	۳۳/۹۷	۵۷/۰۴	۴۲/۵۸	۲۳/۰۶	۳۹/۹۳	۲۹/۲۴
	میانگین تطابق[2]	۳۴/۳۱	۶۰/۹۱	۴۳/۸۹	۲۳/۸۴	۴۲/۴۱	۳۰/۵۲
	نرمال تطابق	۳۴/۹۲	۵۷/۹۳	۴۲/۹۶	۲۳/۰۲	۴۳/۹۳	۳۰/۲۱
خوشه‌بندی	میانگین ورد۲وک	۳۴/۴۵	۵۹/۸۸	۴۳/۷۳	۲۲/۹۸	۴۱/۱۴	۲۹/۴۹
	میانگین وزن‌دار ورد۲وک	۳۴/۸۱	۵۹/۶۵	۴۳/۹۶	۲۳/۲۵	۴۱/۰۳	۲۹/۶۸
	میانگین تطابق	۳۵/۷۹	۶۲/۲۵	۴۵/۴۵	۲۴/۲۵	۴۴/۶۲	۳۱/۴۲
	نرمال تطابق	۳۴/۵۲	۶۳/۶۷	۴۶/۴۲	۲۳/۸	۴۴/۳۹	۳۰/۹۹
ایجاز: امتیازدهی		۳۵/۵۱	۶۰/۱۵	۴۴/۶۶			
فارسی سام: امتیازدهی		۳۸/۶۲	۵۵/۱۲	۴۵/۵۲			

در ادامه، عملکرد الگوریتم پیشنهادی کدگذاری تُنک با استفاده از روش‌های بازنمایی گفته‌شده بر روی دادگان پاسخ، در جدول ۳، آورده شده است. شایان ذکر است که با توجه به سایر پژوهش‌های انجام‌شده بر روی دادگان پاسخ، در این جا نیز طول خلاصهٔ نهایی برابر با ۲۵۰ کلمه در نظر گرفته خواهد شد.

۱. علت وجود خانه‌های خالی عدم ارائهٔ نتایج در مقالات مرجع بر روی سیستم‌های «ایجاز» و «فارسی‌سام» با معیار روژ-۲ است.

2. matching

جدول ۳. نتایج روش پیشنهادی

		روژ-۱		روژ-۲	
نوع مدل	روش بازنمایی	دقت	فراخوانی معیار F	دقت	فراخوانی معیار F
کدگذاری تُنُک	بسامد کلمات در متن و بسامد معکوس مستندات	۵۱/۴۴	۶۸/۱۲ ۵۵/۵۶	۳۵/۱۵	۴۹/۷۲ ۳۸/۰۸
	میانگین ورد۲وک	۵۰/۳۰	۷۲/۶۰ ۵۶/۴۴	۳۵/۴۰	۵۵/۳۹ ۴۰/۰۷

همان‌طور که در نتایج به‌دست آمده مشاهده می‌شود، روش پیشنهادی با تفاوت بالایی بهتر از تمامی مدل‌های پایه قابلیت استخراج خلاصه از متن را دارد. معیار F در هر دو روش ارزیابی روژ-۱ و روژ-۲ بهتر از مدل‌های پایه عمل نموده است. با مقایسهٔ میزان فراخوانی مدل پیشنهادی با مدل‌های پایه مشاهده می‌شود که این مقدار، بهبودی معادل ۵ درصد در هر دو معیار روژ-۱ و روژ-۲ داشته است، ولی میزان بهبود معیار دقت در مدل پیشنهادی به‌مراتب بیشتر و در حدود ۱۵ درصد در روژ-۱ و ۱۱ درصد در روژ-۲ بوده است. از این نتیجه می‌توان دریافت که جملات استخراج‌شده در روش پیشنهادی نسبت به روش‌های پایه صحیح‌تر هستند. این امر می‌تواند تأکیدی باشد بر ارجحیت این روش با در نظر گرفتن اولویت‌های پوشش، تُنُک بودن و تنوع نسبت به سایر روش‌های خلاصه‌سازی. نتیجه این‌که روش ارائه شده در هر دو معیار دقت و فراخوانی توانسته است عملکرد مناسبی داشته باشد.

در گام دیگر، عملکرد روش کدگذاری تُنُک با دو روش بازنمایی مختلف مورد بررسی قرار گرفته است. با مقایسهٔ نتایج حاصل از بازنمایی میانگین «ورد۲وک» و بسامد واژه- معکوس بسامد سند مشاهده می‌شود که بازنمایی میانگین «ورد۲وک» در هر دو معیار روژ-۱ و روژ-۲ توانسته است به نتیجه‌ای بهتر دست یابد. این بهبود همچنین نشان می‌دهد که در نظر گرفتن روابط معنایی کلمات در بازنمایی جملات می‌تواند اثری مثبت در خروجی سیستم و خلاصهٔ استخراج‌شده داشته باشد.

۵. نتیجه‌گیری و پیشنهادات

در چند سال اخیر پژوهش‌های مختلفی برای خلاصه‌سازی فارسی صورت گرفته است. اغلب این تلاش‌ها به پیشرفت چشمگیری در نتایج منتهی نشده و صرفاً پیرامون استفاده و استخراج ویژگی‌های آماری مختلف از متون برای خلاصه‌سازی آن‌ها بوده

است. در این‌جا با استفاده از رویکرد کدگذاری تُنُک و روش‌های نوین بازنمایی جملات سعی بر این بوده است که جملات به بهترین شکل ممکن با حفظ روابط معنایی و ساختاری بازنمایی شوند. در ادامه نیز با استفاده از روش کدگذاری تُنُک، جملاتی که می‌توانند کل متن ورودی را به بهترین شکل ممکن بازسازی نمایند، انتخاب شوند. با ارزیابی روش پیشنهادی بر روی مجموعهٔ دادگان پاسخ نشان داده شد که این روش می‌تواند نسبت به کارهای گذشته نتیجهٔ بهتری داشته باشد. با بررسی نتایج می‌توان دریافت که دلیل بهبود چشمگیر عملکرد، استفاده از الگوریتم کدگذاری تُنُک برای انتخاب جملات خلاصه است. همچنین، استفاده از روش میانگین «ورد۲وک» برای بازنمایی جملات باعث گردید که روابط معنایی میان کلمات جمله به‌شکل بهتری نشان داده شود. از این رو، می‌توان گفت که به‌کارگیری این عوامل نتیجه‌ای مثبت در بهبود خلاصه‌سازی متون ایجاد می‌کند.

برای کارهای آتی پیشنهاد می‌شود که از روش‌های کدگذاری تُنُک دیگری مانند یادگیری گروهی تُنُک استفاده شود. همچنین، با استفاده از مدل‌های بهتر و غنی‌تر بازنمایی جملات می‌توان جنبه‌های مختلف اسناد ورودی را بازنمایی نمود که در نتیجه، الگوریتم انتخاب جملات خلاصهٔ نهایی عملکرد بهتری خواهد داشت. از آنجا که در این پژوهش هدف نشان دادن نقش مؤثر الگوریتم کدگذاری تُنُک و بازنمایی‌های معنایی پایه در بهبود خلاصه‌سازی فارسی بود، بازنمایی‌های پیشرفته‌تر و مبتنی بر بافت مورد بررسی قرار نگرفته است، اما استفاده از این بازنمایی‌ها جهت بهبودهای بیشتر خلاصه‌سازی از جمله کارهای آتی این پژوهش است. با توجه به تلاشی که در این حوزه می‌شود و نیازی که وجود دارد، امید است که روش‌های بهتری برای بازنمایی واحدهای متنی در آیندهٔ نزدیک ارائه گردد.

فهرست منابع

کهنسال، محمود. هشام فیلی، و سعید فرضی. ۱۳۹۶. سامانهٔ خودکار خلاصه‌سازی با استفاده از روش تعبیهٔ متن. *مجموعه مقالات چهارمین همایش ملی زبان‌شناسی رایانشی*. تهران: نشر نویسه پارسی (ص. ۱۶۵-۱۸۶).

References

Behmadi Moghaddas, B., M. Kahani, A. Toosi, A. Pourmasoumi Hassankiadeh, & A. Estiry.2013 . Pasokh: a Standard Corpus for the Evaluation of Persian Text Summarizers 3rd *International*

eConference on Computer .IEEE. Ferdowsi University of Mashhad. Mashhad, Iran.

Gholamrezazadeh, S., M. AminiSalehi, & B. Gholamzadeh. 2009. A Comprehensive Survey on Text Summarization Systems. 2nd International Conference on Computer Science and its Applications. IEEE. Jeju, Korea (South).

Hassel, M & N. Mazdak. 2004. FarsiSum: a Persian text summarizer .Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages. Geneva, Switzerland.

Honarpisheh, M., G. Ghassem-Sani & G. Mirroshandel. 2008. A multi-document multi-lingual automatic summarization system .Proceedings of the Third International Joint Conference on Natural Language Processing. Hyderabad, India.

Hosseinikhah, T., A. Ahmadi A & A. Mohebi. 2018. A new Persian Text Summarization Approach based on Natural Language Processing and Graph Similarity .*Iranian Journal of Information Processing and Management* 33 (2): 885-914.

Karimi, Z & M. Shamsfard. 2006. Summarization of Persian texts .Proceedings of 11th International CSI computer Conference. Tehran, Iran.

Khademi, M., M. Fakhredanesh, & M. Hoseini. 2017. Conceptual Text Summarizer: a new model in continuous vector space. . *Journal of Information Systems and Telecommunications* 7 (1): 23-33.

Kiyoumarsi, F. & F. Rahimi Esfahani. 2011. Optimizing Persian Text Summarization Based on Fuzzy Logic Organization. International Conference on Intelligent Building and Management. Sydney, Australia.

Lin, C.-Y. 2004. Rouge: a package for automatic evaluation of summaries .*Text Summarization Branches Out*. Association for Computational Linguistics. Barcelona, Spain. 84-81.

Liu, H., H. Yu, & Z-H Deng. 2015. Multi-Document Summarization Based on Two-Level Sparse Representation Model .Twenty-ninth AAAI conference on artificial intelligence. Austin, Texas, USA.

Mao, X., H. Yang, S. Huang, Y. Liu. & R. Li. 2019. Extractive summarization using supervised and unsupervised learing. *Expert Systems with Applications* 133: 173-181.

Masoumi, S., M-R Feizi-Derakhshi, & R. Tabatabaei.2014. TabSum- a new Persian text summarizer.. *Journal of mathematics and computer science* 11 (4): 330-342.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, & J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13). Curran Associates Inc., Red Hook, NY, USA. 3111–3119.

Parvandeh, S., S. Lahiri, & F. Boroumand. 2016. PerSum: Novel Systems for Document Summarization in Persian. *International Journal of Asian Language Processing* 26 (2): 67-108.

Pour-masoomi, A., M. Kahani, S. A. Toosi, & A. Estiri. 2014. Ijaz: an Operational system for single-document summarization of Persian news texts. *Signal and Data Processing* 11 (1): 33-48.

Rohanian, M. 2017. Multi-Document Summarization of Persian Text Using Paragraph Vectors. *Proceedings of the Student Research Workshop associated with RANLP*. Varna, Bulgaria.

Shakeri, H., S. Gholamrezazadeh, M. Salehi, & F. Ghadamyari. 2012. A New Graph-Based Algorithm for Persian Text Summarization .*Computer science and convergence* . Dordrecht: Springer.

Shamsfard, M., T. Akhavan, & M. Erfani Jourabchi. 2009. Parsumist: a Persian text summarizer. International Conference on Natural Language Processing and Knowledge Engineering .IEEE. Dalian, China.

Soltani, M., J. Nasiri, & E. Asgarian. 2018. An Automatic Persian Text Summarization System Based on Linguistic Features and Regression .*Iranian Journal of Information Processing and Management* 33 (4): 1809-1828.

Tofighy, M., O. Kashefi, A. Zamanifar, & H. Haj Seyyed Javadi. 2011. Persian Text Summarization

Using Fractal Theory. Informatics Engineering and Information Science. Berlin Heidelberg. Berlin, Heidelberg: Springer 651-662.

Tofighy, M., R. G. Raj, & H. Haj Seyyed Javadi. 2013. APH Techniques for Persian Text Summarization. *Malaysian Journal of Computer Science* 26 (1): ?

رامین فتوره‌چی

دارای مدرک کارشناسی ارشد در رشتهٔ کامپیوتر، گرایش هوش مصنوعی از دانشکدهٔ مهندسی کامپیوتر دانشگاه صنعتی امیرکبیر است. مباحث پردازش متن، داده‌کاوی و مهندسی نرم‌افزار از جمله علایق پژوهشی وی است.

سعیده ممتازی

دارای مدرک دکتری هوش مصنوعی از دانشگاه زارلند آلمان است. ایشان هم‌اکنون استادیار دانشکدهٔ مهندسی کامپیوتر دانشگاه صنعتی امیرکبیر است. وی سرپرستی آزمایشگاه پردازش زبان طبیعی و مرکز تحقیقات فناوری اطلاعات و ارتباطات دانشگاه صنعتی امیرکبیر را بر عهده دارد.

پردازش زبان طبیعی از جمله علاقه پژوهشی وی است.