مجله‌ی علمی پژوهش در علوم رایانه، شماره ۲۸، زمستان ۱۴۰۱، ص ۴۷- ۳۳

ISSN: 2745-3673
http://www.Rcsj.ir

مجله علمی پژوهش در علوم رایانه (سال هفتم)
شماره ۲۸ / زمستان ۱۴۰۱ / ص ۴۷- ۳۳

ارائه یک سیستم خلاصه ساز کاربردی بر مبنای الگوریتم رتبه صفحه و الگوریتم‌های فراابتکاری برای زبان فارسی

احسان باقری[۱]، مهدی حسنی نسب[۲]، ایوب محمدیان[۳]

[۱] کارشناسی ارشد، مدیریت فناوری اطلاعات موسسه غیرانتفاعی نورطوبی.
[۲] دکترا، دانشکده علوم و فناوری های نوین، دانشگاه تهران، تهران، ایران.
[۳] استادیار، سرپرست مرکز نوآوری کسب و کار دانشکده مدیریت معاونت آموزشی دانشکده مدیریت.

نام نویسنده مسئول:
احسان باقری

تاریخ دریافت: ۱۴۰۱/۰۹/۳۱
تاریخ پذیرش: ۱۴۰۱/۱۲/۰۸

چکیده

خلاصه سازی خودکار متن، یک پژوهش ضروری در پردازش زبان طبیعی است که تلاش می‌کند اسناد متنی را خلاصه کند تا کاربران بتوانند به سرعت به اطلاعات مفید دسترسی پیدا کنند. با وجود اینکه در زبان فارسی تلاش‌هایی برای ایجاد خلاصه‌سازی متون صورت گرفته است، اغلب موارد این پژوهش‌ها به فرم نظریه ارائه شده است. در این پژوهش سعی میشود یک سیستم بر اساس رتبه متن و گراف تحلیلگر نحوی جملات عمل کند و سپس نتایج برای پردازش به الگوریتم معروف رتبه صفحه هدایت شود که جملات با بالاترین رتبه گراف را استخراج کند. سپس به همین روش، کلمات استخراج میشود و نهایتا با استفاده از الگوریتم‌های فرااکتشافی، برخی از جملات دارای کلمات کلیدی به متن استخراج شده افزوده میشوند. لازم به ذکر است اگرچه الگوریتم پیشنهادی، از تحلیل‌صرف و نحوی برای شباهت بین جملات استفاده میکند، با این‌حال این کار در زبان فارسی میتواند توسط الگوریتم‌های مختلفی صورت پذیرد. با توجه به اینکه تاکنون سیستم مذکور پیشنهاد نشده است و همچنین در خصوص مقایسه تحلیل نحوی جملات در زبان فارسی نیز تحقیق جامعی انجام نگرفته است، این موضوعات جدید محسوب میشود. نتایج برنامه نه تنها به با جامعه آماری وسیع تری نسبت به موارد مشابه مقایسه گردیده است بلکه از نظر متون تخصصی و استفاده از سخت افزار معمول هم مورد توجه قرار گرفته است و در همه حالات توانسته نتایج عملی قابل قبولی را کسب نماید.

واژگان کلیدی: خلاصه سازی تک سندی، پردازش زبان، تحلیل صرف ونحوی، الگوریتم رتبه صفحه.

ISSN: 2745-3673
http://www.Rcsj.ir

مقدمه

در عصر دیجیتال امروز، تغییرات در عرصه‌های مختلف به گونه‌ای است که مبنای رقابت در اقتصاد ملی و جهانی، از منابع مشهود مانند نفت، طلا و غیره به منابع غیر مشهود، مانند اطلاعات و توانایی پردازش آن تغییر یافته است. لذا، عصر حاضر را عصر اطلاعات و نهادهای مبتنی بر دانش می‌دانند.

یکی از ویژگی‌های اساسی عصر دیجیتال، سرعت است. در این عصر، پیشرفت و تحول در کمترین زمان ممکن اتفاق می‌افتد. کسب وکارها با راهبردهای مناسب به سرعت رشد می‌کنند و با استراتژی‌های نادرست به سرعت از صحنه رقابت حذف خواهند شد[1]. بانک جهانی، داده را به‌عنوان سوخت اقتصاد دیجیتال معرفی می‌کند. داده به سازمان‌های این عصر کمک می‌کند تا از فضای حدس و گمان خارج شده و به سوی پیش‌بینی‌های الهام بخش و آزمودن مستمر فرضیه‌ها حرکت کنند. امروز داده‌ها بینش لازم برای تصمیم‌گیری را فراهم می‌آورند و دوران تصمیم‌گیری صرف برپایه شهود و احساسات به پایان رسیده است. کسب‌وکارهای امروز نیز، به کسب‌وکارهای مبتنی بر تحلیل داده تبدیل شده‌اند. آن‌ها برای دستیابی به شناخت بیشتر از مشتریان، رقبا، بازارهای هدف و حتی محصولات و خدمات قابل ارائه، نیازمند تحلیل داده و اطلاعات هستند [2]. با توجه به این موضوع که اکثر گزارش‌ها و اسناد در قالب متن هستند میتوان گفت اگر کاربران مجهز به سیستم خلاصه‌ساز باشند این امکان را خواهند داشت تا دسترسی سریعتر به اطلاعات ورودی بدون نیاز به خواندن تمام آن‌ها داشته باشند. این باعث ایجاد منابع بیشتر و با سرعت بالاتر و حاصل شدن اطلاعات غنی‌تر می‌گردد.

معرفی مسئله

برای زبان دو رویکرد اصلی خلاصه سازی متن وجود دارد که عبارتند از: استخراجی و انتزاعی. راه حل استخراجی مستلزم انتخاب جملات خاص از بدنه متن برای ایجاد خلاصه نهایی است. رویکرد کلی به راه‌حل‌های استخراجی با رتبه‌بندی جملات بر اساس اهمیت آنها در متن و برگرداندن مهم‌ترین جملات به کاربر همراه است. با این حال، راه‌حل‌های انتزاعی برای خلاصه‌سازی متن شامل ایجاد جملات جدید برای به تصویر کشیدن متن و معنا در پشت بدنه اصلی متن است. اگرچه انسان ها به این شکل به خلاصه سازی متن می پردازند، اما تعمیم و آموزش آن به یک ماشین کار بسیار دشواری است. تکنیک‌های فشرده سازی متن معمولاً برای حل رویکردهای انتزاعی در خلاصه‌سازی متن استفاده می شود. عوامل مختلفی در خلاصه‌سازی متن وجود دارد که تأثیری را که ممکن است بر داستان اصلی بگذارد، تغییر می‌دهد. به این جهت اکثرا خلاصه معنی‌داری تولید نمیشود. برای حل مسئله میتوان به دو استراتژی شناخته شده اشاره کرد: اول راهکار استخراجی استفاده از الگوریتم TextRank است. روش بعدی انتزاعی است و استفاده از شبکه عصبی رمزگذار-رمزگشای Seq2Seq[1] میباشد و استراتژی سوم که ممکن است به دلیل شباهت‌های زیاد در استرتژی دوم هم قرارگیرد استفاده از معماری شایع Transformers است که هریک از این موارد کاربردهای مختلفی دارند. اینجا توسعه و پیاده‌سازی یک خلاصه‌ساز با روش استخراج متن برای کاربردهای مختلف مورد توجه میباشد. انتظار میرود که برنامه توسعه داده شده روی سیستمهای معمولی قابل اجرا و کاربردی باشد.

کارهای مرتبط

اگرچه بسیاری از روش‌های خلاصه‌سازی متن برای زبان‌هایی مانند انگلیسی در دسترس است، اما پژوهش‌های انجام شده برای فارسی اغلب از جنس نظریه پردازی هستند. این روش‌ها را می‌توان به دو دسته تحت نظارت و بدون نظارت تقسیم کرد.روش‌های خلاصه‌سازی نظارت شده برای اسناد فارسی به چهار دسته روش‌های اکتشافی، مبتنی بر زنجیره واژگانی، مبتنی بر نمودار، و یادگیری ماشینی یا روش‌های مبتنی بر ریاضی تقسیم می‌شوند.

[1] Encoder-Decoder Seq2Seq

ISSN: 2745-3673
http://www.Rcsj.ir

روش اکتشافی

این این روش به تجزیه و تحلیل داده های اکتشافی روی داده ها به منظور کشف الگوها، تشخیص ناهنجاری ها، آزمایش فرضیه ها و بررسی فرضیات با کمک آمار اشاره دارد.

هاسل و مزدک FarsiSum را به عنوان یک روش اکتشافی پیشنهاد کردند. این یکی از اولین تلاش ها برای ایجاد یک سیستم خلاصه سازی خودکار متن برای فارسی است. این سیستم به عنوان یک برنامه با زبان پرل[۲] نوشته شده و از ماژول های پیاده سازی شده در [3] SweSum استفاده شده است که مربوط به زبان سوئدی میباشد، برنامه از یک لیست توقف فارسی در قالب یونیکد و مجموعه کوچکی از قوانین اکتشافی استفاده کرده است[4].

زمانی‌فر و همکاران. [5] یک تکنیک خلاصه ترکیبی جدید را پیشنهاد کردند که از اصطلاح خصوصیت همزمانی[۳] متن استفاده میکند. این تکنیک می تواند تعداد سوژه‌ها را تشخیص دهد. تکنیک پیشنهادی، سند را متناسب با موضوع مورد بررسی در یک سند خلاصه می‌کند .ویژگی مفهومی الگوریتم متن را در نظر می گیرد و بر اساس مترادف کلمه از گنجاندن جملات مشابه در خلاصه جلوگیری می کند. آنها در مقایسه با FarsiSum عملکرد بهتری را در آزمون توسط اشخاص گزارش میکنند.

شمس‌فرد و همکاران تکنیک Parsumist را ارائه داده‌اند. آنها روش‌های خلاصه‌سازی تک سندی و چند سندی را با استفاده از زنجیره‌های واژگانی و نمودارها ارائه کردند.که از ترکیبی از آماری، معنایی و اکتشافی بهبود یافته بهره‌برداری می کند برای رتبه بندی و تعیین مهم ترین جمله، بیشترین شباهت را با جملات دیگر، عنوان و کلیدواژه‌ها در نظر می‌گیرند. آنها عملکرد بهتری نسبت به FarsiSum گزارش کردند [6].

زمانی فر و کاشفی روش AZOM را پیشنهاد کردند. این روش ویژگی‌های متن آماری و مفهومی را ترکیب می کند و از نظر ساختار سند، خلاصه متن را استخراج می کند. همچنین قادر است اسناد بدون ساختار را نیز خلاصه کند. رویکرد پیشنهادی این پژوهش برای زبان فارسی نیز بومی‌سازی شده است[7].

شفیعی و شمس‌فرد یک خلاصه‌کننده تک/چند سندی را با استفاده از روش خوشه‌بندی جدید برای ایجاد خلاصه‌ها ارائه کردند. اساس کار این روش به این صورت است که در ابتدا یک مرحله انتخاب ویژگی استفاده می شود، سپس از فارس‌نت و یا همان ورد نت فارسی، برای استخراج اطلاعات معنایی کلمات استفاده می شود. بنابراین، جملات ورودی به سه خوشه اصلی تقسیم می‌شوند: یعنی سه کلاستر تشکیل میشود و هر یک یک فضای خاص را بررسی میکند. کلاستر اول شباهت، کلاستر دوم ارتباط و کلاستر سوم انسجام هر کلاستر نوع اول - شباهت حاوی جملات مشابه با هسته خود است. هر کلاستر نوع دوم- ارتباط حاوی جملاتی است که به هسته خود مرتبط هستند (اما نه مشابه). کلاستر های انسجام جملاتی را نشان می دهند که برای حفظ انسجام خلاصه باید کنار هم قرار گیرند. در نهایت، مرکز هر کلاستر شباهت دارای بیشترین امتیاز ویژگی به یک خلاصه خالی اضافه می شود. خلاصه با گنجاندن جملات مرتبط از کلاستر های مرتبط و حذف جملات مشابه به محتوای آن به صورت تکراری بزرگ‌تر می‌شود. کلاستر های انسجام در آخرین مرحله به خلاصه ایجاد شده اعمال می شوند [8].

حسینی خواه و همکاران یک روش استخراجی را با ترکیب تکنیک های پردازش زبان طبیعی و متن کاوی پیشنهاد کردند. بخشی از برچسب گذاری گفتار برای محاسبه ضریب اهمیت کلمات و از روش شباهت نمودار برای انتخاب جملات بدون مشکل افزونگی (تکرار موارد نزدیک به هم) استفاده می شود [9].

روش های مبتنی بر ریاضی و یادگیری ماشینی

این روش به استفاده از الگوریتم های ریاضی بخصوص احتمالات و انواع مختلف الگوریتم های یادیگری ماشین اشاره دارد.

کیومرثی و رحیمی [10] روش جدیدی را برای خلاصه کردن متون فارسی بر اساس ویژگی های موجود در زبان فارسی و استفاده از منطق فازی پیشنهاد کردند مبنای کار بر اساس یک پیشنهاد انعطاف پذیر جملات میباشد که با استفاده از ۱۲ ویژگی

[2] Perl
[3] Co-Occurrence

ISSN: 2745-3673
http://www.Rcsj.ir

جملات شروع به امتیاز دهی جملات میکند. این خصوصیات شامل مواردی مانند طول جمله، شباهت به عنوان، شباهت به کلمات کلیدی و غیره میباشند.

توفیقی و همکاران [11] روش جدیدی را برای خلاصه‌نویسی متن فارسی مبتنی بر نظریه فراکتال پیشنهاد کرد که هدف اصلی آن استفاده از ساختار سلسله‌مراتبی سند برای بهبود کیفیت خلاصه‌سازی متون فارسی است. مطابق نتایج حاصل از این تحقیق عملکرد بهتری نسبت به FarsiSum و ضعیف تر از AZOM گزارش شد

بازغندی و همکاران [12] یک سیستم خلاصه سازی متنی را بر اساس خوشه بندی جملات پیشنهاد کرد. برای بهینه سازی روش ها از الگوریتم های هوش جمعی استفاده شده‌است. این روش ها بر اساس روابط آنها در متن بر جنبه معنایی کلمات تکیه دارند. نتایج آنها با رویکردهای خوشه بندی سنتی قابل مقایسه است.

توفیقی و همکاران [13] یک تکنیک فرآیند تحلیل سلسله مراتبی[۴] را برای خلاصه سازی متن فارسی پیشنهاد کرد. مدل پیشنهادی از سلسله مراتب تحلیلی به عنوان عامل پایه برای الگوریتم ارزیابی استفاده می کند.مطابق نتایج تحقیق صورت گرفته درمقایسه با FarsiSum عملکرد بهتری گزارش شده است.

پورمعصومی و همکاران [14] یک سیستم خلاصه سازی تک سندی فارسی به نام ایجاز را پیشنهاد کرد. بر اساس روش حداقل مربعات وزنی[15] است. آنها در مقایسه با FarsiSum عملکرد بهتری را گزارش کردند. آنها همچنین مجموعه ای برای ارزیابی خلاصه نویسان متن فارسی به نام پاسخ را پیشنهاد کردند [16].

فرضی و کیانیان [17] خلاصه‌نویس فارسی را بر اساس رویکرد خلاصه‌سازی نیمه‌نظارت[۵] شده کتیبه را پیشنهاد کردند که ترکیبی از الگوریتم‌های هم‌آموزشی[۶] و خودآموزی[۷] است. آموزش ترکیبی[۸]، زمانی استفاده می‌شود که فقط مقادیر کمی از داده‌های برچسب‌دار و مقادیر زیادی داده بدون برچسب وجود داشته باشد. در این پژوهش یک رویکرد نیمه نظارتی برای غلبه بر فقدان داده‌های برچسب‌دار کافی ارائه شده‌است.

فخردانش و همکاران با استفاده از ترکیبی از یادگیری عمیق[۹] و روش های آماری، مفاهیم متن را خوشه بندی کرده و بر اساس اهمیت مفاهیم ، جمله با بیشترین بار مفهومی را برگزیده است. در این روش بدون نظارت[۱۰]و بدون استفاده از ویژگی‌های دست‌ساز، نتایج پیشرفته‌ای را در مجموعه اسنادی پاسخ در مقایسه با بهترین روش‌های فارسی نظارت شده را گزارش کرده اند. اساس کار مدل کیسه کلمات[۱۱] به هم پیوسته میباشد و از پیکره[۱۲] همشهری استفاده شده است [18].

استفاده از الگوریتم‌های فرابتکاری

هدف امتیازدهی به جملات، با تأکید بر برخورد منصفانه با ویژگی های متن بر اساس اهمیت کلمات کلیدی است. در طول دهه گذشته، اشکال مختلفی از الگوریتم فرا ابتکاری حل مسئله خلاصه سازی متن را همراهی کرده است. مثلا یکی از این فرا ابتکاری‌ها، بهینه سازی ازدحام ذرات [۱۳]است که رفتار یک گروه ماهی یا گله پرندگان را شبیه سازی می کند. الگوریتم کلنی مورچگان[۱۴] که از جامعه مورچه ها الهام گرفته شده است. و الگوریتم کلونی زنبورهای مصنوعی[۱۵] رفتار زنبورهای عسل را شبیه سازی می کند. اخیراً، فرا ابتکاریهای جدید مانند جستجوی فاخته [۱۶]نیز در زمینه خلاصه سازی استفاده شده است.

---

[4] AHP
[5] Semi supervised
[6] co-training
[7] self-training
[8] combination
[9] Deep learning
[10] unsupervised
[11] Bag of words
[12] corpus
[13] PSO
[14] ACO
[15] ABC
[16] CS

ISSN: 2745-3673
http://www.Rcsj.ir

جدول ۱- تاریخچه برخی الگوریتمهای فراابتکاری برای خلاصه سازی متن

الگوریتم	سال انتشار
مبتنی بر ازدحام ذرات	۲۰۰۹[19] - ۲۰۱۵ [20]-۲۰۲۱ [21]-۲۰۲۲ [22]
مبتنی بر کرم شب تاب	۲۰۲۲ [23] - ۲۰۱۹ [24]
مبتنی بر کلنی مورچه	۲۰۱۷ [25]
مبتنی بر کلنی زنبور عسل	۲۰۱۷ [26]

جدول۲- مرور ادبیات برای خلاصه ساز

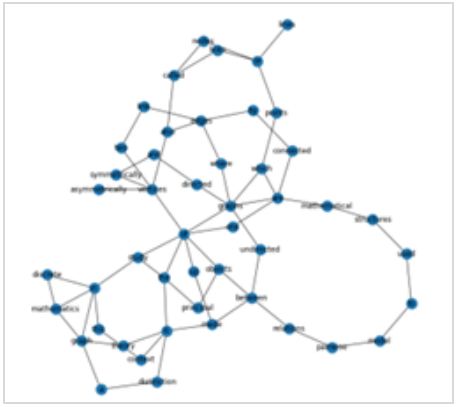
نوع	سال	ماجول	توضیحات
اکتشافی	۲۰۰۴	FarsiSum	از SweSum [3] استفاده شده است که مربوط به زبان سوئدی میباشد به عنوان یک روش اکتشافی پیشنهاد شده [4].
	۲۰۰۹	Co-Occurrence and Conceptual	یک تکنیک خلاصه ترکیبی جدید را پیشنهاد شد که از اصطلاح خصوصیت همزمانی متن استفاده میکند. [5].
	۲۰۰۹	Parsumist	آنها روش‌های خلاصه‌سازی تک سندی و چند سندی را با استفاده از زنجیره‌های واژگانی و نمودارها ارائه کردند. [6].
	۲۰۱۱	AZOM	ویژگی های متن آماری و مفهومی را ترکیب می کند و از نظر ساختار سند، خلاصه متن را استخراج می کند [7].
	۲۰۱۷	Similarity versus relatedness	استفاده از روش خوشه‌بندی جدید برای ایجاد خلاصه‌ها ارائه می‌کند. ابتدا یک مرحله انتخاب ویژگی استفاده می شود. [8].
	۲۰۱۷	Natural Language Processing and Graph Similarity	ترکیب تکنیک های پردازش زبان طبیعی و متن کاوی [9].
یادگیری ماشین	۲۰۱۱	Fuzzy Logic Approach	یک روش اساس ویژگی های موجود در زبان فارسی و استفاده از منطق فازی [10].
	۲۰۱۱	Fractal Theory	اساس نظریه فراکتال پیشنهاد کرد که هدف اصلی آن استفاده از ساختار سلسله مراتبی سند برای بهبود کیفیت خلاصه‌سازی متون [11]
	۲۰۱۲	Based On PSO Clustering	خوشه بندی جملات و هوش جمعی [12]
	۲۰۱۲	AHP	تحلیل سلسله مراتبی (AHP) را برای خلاصه سازی متن فارسی پیشنهاد کرد [13].
	۲۰۱۴	Ijaz	رویکرد خلاصه‌سازی نیمه نظارت شده [14].
	۲۰۱۸	Katibeh	ترکیبی از یادگیری عمیق و روش های آماری، مفاهیم متن را خوشه بندی و بر اساس اهمیت مفاهیم هر جمله، با بیشترین بار مفهومی. در روش بدون نظارت، بدون استفاده از ویژگی‌های دست‌ساز [17].
	۲۰۱۹	Continuous Vector Space	ترکیبی از یادگیری عمیق و روش های آماری، مفاهیم متن را خوشه بندی و بر اساس اهمیت مفاهیم هر جمله [18].

با توجه به اینکه تعداد محدودی از سورسها و پیکره های فوق در دسترس هستند ابراز نظر قطعی در مورد کارکرد آنها دشوار است. پیکره پاسخ در حال حاضر عملا در دسترس نیست و نمیتوان ارزیابی صحت بر روی آن انجام داد. از طرفی،در بسیاری موارد راهکارهای عملی در مورد پیاده سازی وجود ندارد.

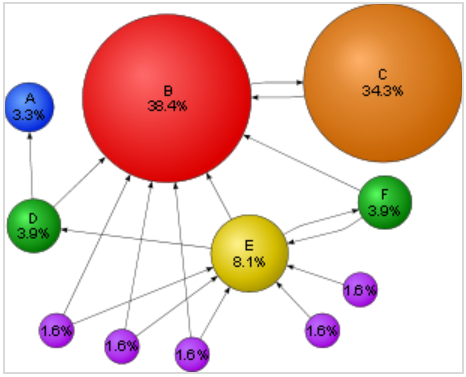
ISSN: 2745-3673
http://www.Rcsj.ir

روش پیشنهادی

در روش پیشنهادی اساس کار به این صورت است که ابتدا یک لیست از کلمات و تحلیل صرف و نحوی فارسی آنها ایجاد میشود سپس گراف تحلیل ایجاد میشود. شیوه ایجاد گراف تحلیل به این صورت است که ابتدا یک لیست یکتا از کلمات همراه با برچسب نحوی آنها ایجاد میشود. خروجی کار یک یونی دایرکت[۱۷] گراف است. سپس الگوریتم PageRank گراف را دریافت میکند طرز کار به این شکل است که ابتدا باید در صورت لزوم گراف را به یک گراف دایرکت[۱۸] تبدیل کرد. سپس یک کپی به شکل تصادفی ایجاد شده، بردار شروع ثابت را که نرمال شده است انتخاب و در نهایت گره های آویزان[۱۹] گراف را با استفاده از یک حلقه دریافت می‌شود.جهت بررسی همگرایی، هنجار ۱۱ را انجام میشود. و در نهایت یک رنک وکتور را ایجاد میکند که برای کلمات کلیدی مناسب است. مجموع گره های آویزان برای جملات استفاده شده و یک رنک وکتور را ایجاد میکند. در نهایت و جملات در آن به ترتیب اهمیت مرتب می‌شوند. با توجه به اندازه جملات و خروجی مورد انتظار تعدادی از جملات را از بالای لیست انتخاب میشود. سپس همین مراحل را برای پیدا کردن کلمات کلیدی طی میشود.



شکل۱- نمایش فرضی گراف یونی دایرکت کلمات



شکل۲- نمایش فرضی تاثیر TextRank

حاصل کلمات کلیدی را که با استفاده از تحلیل نحوی پیدا شده را به الگوریتم کلنی مورچه داده و نهایتا ترکیب جملات استخراج شده و جملات با کلمات با کلیدی موثر مشخص میشود.

تحلیل متنی

به طور کلی چهار روش را میتوان برای این موضوع در نظر گرفت:

[17] undirected graph
[18] directed graph
[19] dangling node

ISSN: 2745-3673
http://www.Rcsj.ir

۱. روش‌های مبتنی بر واژگان - تگ POS را که اغلب با یک کلمه در مجموعه آموزشی اتفاق می‌افتد، اختصاص می‌دهد.
۲. روش‌های مبتنی بر قانون - برچسب‌های POS را بر اساس قوانین اختصاص می‌دهد. برای مثال، می‌توان قاعده‌ای داشت که می‌گوید، کلماتی که به «ed» یا «ing» ختم می‌شوند باید به یک فعل اختصاص داده شوند. تکنیک‌های مبتنی بر قانون را می‌توان همراه با رویکردهای مبتنی بر واژگان استفاده کرد تا امکان برچسب‌گذاری POS کلماتی را که در مجموعه آموزشی وجود ندارند، اما در داده‌های آزمایشی وجود دارند، فراهم کند.
۳. روش‌های احتمالی- این روش تگ‌های POS را بر اساس احتمال وقوع یک دنباله برچسب خاص اختصاص می‌دهد. میدان‌های تصادفی شرطی[20] و مدل‌های مارکوف پنهان[21] رویکردهای احتمالی برای اختصاص تگ POS هستند [27] [28].
٤. روش‌های یادگیری عمیق - شبکه‌های عصبی مکرر همچنین می‌توانند برای برچسب‌گذاری POS استفاده شوند[29] . هدف پیدا کردن الگوریتم با دقت بالا و هزینه پایین است.

در ادامه چند روش برای زبان فارسی و پیکره بی جن خان تحلیل و آزمایش شده است.

جدول ۳- مقایسه روشهای تحلیل صرف و نحوی با پیکره بی جن خان

روش تحلیل جمله	میزان دقت	استفاده از منابع سیستمی
Naive Bayes	۷۵.۶۵	متوسط
درخت تصمیم	۹۶.۵۵	متوسط
حداکثر آنتروپی	۹۵.۸۹	کم
مدل مخفی مارکوف	۹۲.۱۵	کم
تحلیل گر استنفورد	۷۹.۸۵	متوسط

تقسیم پاراگراف به جملات

برای تقسیم پاراگراف هم میتوان از روشهای مختلفی استفاده کرد. در برنامه ایجاد شده از OpenNLP استفاده شده است. کلاس MaximumEntropySentenceDetector میتواند آموزش داده شود تا با جملات فارسی سازگاری داشته باشد. به طور خلاصه آشکارساز جمله OpenNLP می تواند تشخیص دهد که یک کاراکتر، نقطه گذاری پایان یک جمله را نشان می دهد یا نه. به این معنا، جمله به عنوان طولانی‌ترین دنباله کاراکتر بریده شده با فاصله سفید بین دو علامت نگارشی تعریف می‌شود. جمله اول و آخر از این قاعده مستثنی است. اولین کاراکتر بدون فاصله شروع یک جمله و آخرین کاراکتر بدون فضای خالی پایان جمله در نظر گرفته می شود.

ایجاد گراف جملات

در این بخش یک لیست از جملات که توسط بخش قبل ایجاد کرده ایم را به تابع میدهیم تا یک یونی گراف بسازد.

ایجاد یونی دایرکت گراف

برای ایجاد این نوع گراف از الگوریتم فاصله لونشتاین استفاده شده است. فاصله لونشتاین عددی است که به شما نشان می دهد که دو رشته چقدر متفاوت هستند. هرچه این عدد بیشتر باشد، تفاوت دو رشته بیشتر است. از نظر شهودی، درک فاصله لونشتاین بسیار آسان است. اساساً دلالت بر این دارد که فاصله خروجی بین این دو مجموع تجمعی تک کاراکتر است

با این تفاسیر ما لبه ها را به نودها اضافه میشود با استفاده از این حالت ما نودهای گراف را دارای همسایه هایی شده است که هر یک وزن مشخصی دارند.

[20] CRF
[21] HMMs

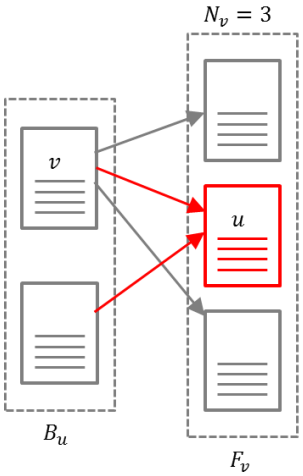
ISSN: 2745-3673
http://www.Rcsj.ir

استفاده از الگوریتم PageRank برای ایجاد فهرست رتبه ها

کاربرد این الگوریتم شناخته شده گوگل تقریبا مشخص است که قبلا در مورد آن توضیحاتی داده شد. این الگوریتم اساساً وزن بین جملات را با مشاهده اینکه کدام کلمات با هم همپوشانی دارند محاسبه می کند. هنگام اعمال به دنبال کلمه مثلا (اسم ها/صفت ها) می گردد که در جملات یکسان هستند و سپس رتبه صفحه گوگل را در شبکه جمله اعمال می کند.

$$R(u) = \sum_{v \in Bu} \frac{R(v)}{Nv}$$

- R(u) برابر رتبه صفحه وب u
- Bu مجموعه ای از صفحات وب که به صفحه وب u ارجاع می دهند
- Fv مجموعه ای از صفحات وب که صفحه وب v به آنها ارجاع می دهد
- Nv تعداد لینک های فوروارد صفحه وب v



شکل ۳- بک لینک های صفحه 'u'

رتبه یک صفحه وب با مجموع تمام رتبه های دریافتی از بک لینک های آن تعیین می شود. و رتبه به طور مساوی با تعداد پیوندهای رو به جلو آن توزیع می شود. انتشار رتبه ای به صورت تکراری انجام می شود. این نه تنها تأثیرات صفحات وب واقع در نزدیکی، بلکه تأثیرات صفحات وب در دوردست را نیز در نظر می گیرد.

$$R_{i+1} = S^T R_i$$

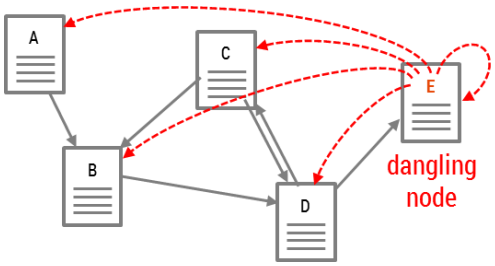
- Ri: بردار رتبه ای در تکرار i
- S: هسته انتقال که در آن مجموع ورودی های یک ردیف ۱ است

دلیل استفاده از ماتریس تصادفی جابجا شده[۲۲]، تنظیم جهت فرآیند است که رتبه یک صفحه وب، که با مجموع رتبه های داده شده از بک لینک ها تعیین می شود.

ماتریس تصادفی دارای یک ماتریس مربع است که مجموع هر سطر آن ۱ است. اما پیوندهای آویزان[۲۳] به سادگی پیوندهایی هستند که به هر صفحه ای بدون لینک خروجی اشاره می کنند. آنها مدل را تحت تأثیر قرار می دهند زیرا مشخص نیست وزن آنها کجا باید توزیع شود و از طرفی تعداد آنها نیز زیاد است.

[22] transposed stochastic matrix
[23] dangling node

ISSN: 2745-3673
http://www.Rcsj.ir



شکل ۴- نودهای اویزان

از آنجایی که پیوندهای آویزان مستقیماً بر رتبه بندی هیچ صفحه دیگری تأثیر نمی گذارد، به سادگی از سیستم حذف می شوند تا زمانی که تمام PageRanks محاسبه شود. پس از محاسبه همه رتبه‌های صفحه، می‌توان آن‌ها را مجدداً اضافه کرد بدون اینکه تأثیر قابل‌توجهی بر محاسبات بگذارد.در نرم افزار ایجاد شده مجموع گره های آویزان را با میزان وزن گره های آویزان ضرب شده در حالیکه که وزن توسط مقدار تقسیم بر مجموع مقدار گره به دست آمده است سپس مقدار ضرب در یک مقدار ثابت (آلفا) شده و در نهایت هم در یک مقدار پیش وزنی ضرب گردیده است. مقدار آلفا ۰.۲۵ فرض شده است. مقدار پیش وزنی هم برابر با وزن درنظر گرفته شده است. نهایتا کلاس PageRank رتبه بندی گره ها در نمودار G را بر اساس ساختار پیوندهای ورودی محاسبه می کند.

```
foreach (GraphNode<T> page in _x.Keys.ToList())
        {
          var nbrs = sGraph[page];
          foreach (var nbr in nbrs)
          {
            var xn = _xLast[page];
            var we = sGraph[page, nbr.GraphNode as GraphNode<T>];
            _x[nbr.GraphNode as GraphNode<T>] += alpha * xn * we;
          }

          var _dnglWeight = _dangling_weights.ContainsKey(page) ? _dangling_weights[page] : 0;
          var _perWeight = _dangling_weights.ContainsKey(page) ? _dangling_weights[page] : 0;
          _x[page] += danglesum * _dnglWeight + (1.0 - alpha) * _perWeight;
        }
```

شکل ۴-۴ کد سی شارپ برای هر اافزودن مقدار به هر نود

مرتب سازی رنک وکتور

در این بخش تنها بر اساس عدد به دست آمده از مرحله قبل جملات را لیست کرده و مرتب میشود

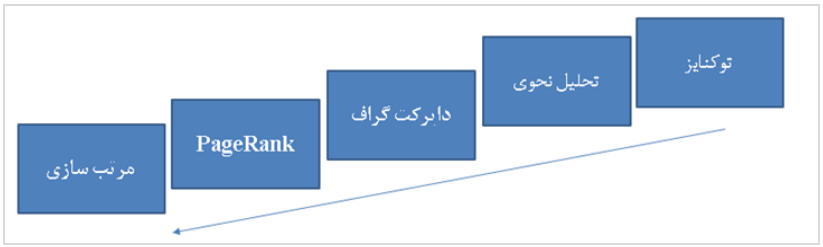
تاثیر کلمات کلیدی

در خصوص خلاصه‌سازی بر اساس کلمات کلیدی مطالعات مختلفی انجام شده‌است. به عنوان مثال هرناندز و همکاران یک سیستم بر اساس خوشه‌ها و کلمات کلیدی خودکار را معرفی کرده اند [۳۰]. یا روشهای دیگری هم با استفاده از یادیگری نیمه نظارت شده وجود دارد [۳۱] در این پژوهش به عنوان یک راه حل بهینه و ترکیبی برای نتیجه از این روش استفاده شده است. .

در این قسمت ابتدا جملات را به طریقی که ذکر شد توکنایز میشود سپس با استفاده از تحلیل نحوی به صورت صرف و نحوی برچسب‌گذاری میگردد. سپس مقادیر را در فهرست‌های کلید و مقدار قرار میگیرد. سپس یک دایرکت گراف تشکیل میشود که از مقادیر یکتای نرمال شده ایجاد شده است. مراحل ایجاد دایرکت گراف مانند بخش قبل میباشد و فاصله لوینشتاین

ISSN: 2745-3673
http://www.Rcsj.ir

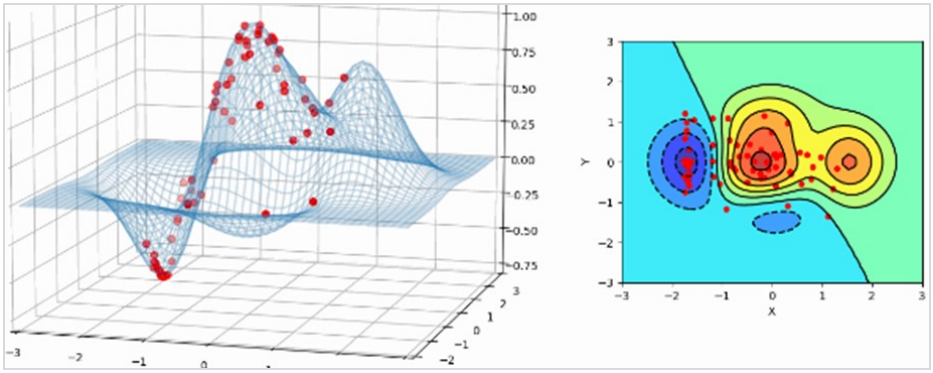
و لبه ها را دارد. سپس با استفاده از PageRank یک فهرست رنک ایجاد میشود که این موارد هم مانند قبل میباشد. خروجی کار یک لیست از کلمات است که مرتب میشود. در این حالت فرض میشود که یک مسیر بین کلمات کلیدی وجود دارد که بر اساس فاصله کلمات از هم در متن شکل گرفته است. اگر مسیر ما با مسیر جستجوی الگوریتم فرابتکاری و پیدا کردن کوتاه ترین سفر یکسان باشد جمله وارد متن میشود و در غیر این صورت از کلمه کلیدی عبور میشود. این بخش در ادامه توضیح داده خواهد شد.



شکل۱-۴- شمای کلی الگوریتم برای کلمات کلیدی

استفاده از الگوریتم جدید

در این مرحله یک روش ابداعی برای بهینه سازی ایجاد شده است. بر این مبنا که کلماتی که کلیدی هستند ممکن است در جملاتی باشند که انتخاب نشده اند بنابر این کلمات کلیدی یک مفهوم پیوسته در متن هستند و احتمالا برای انتقال معنی نویسنده آن استفاده میشوند.

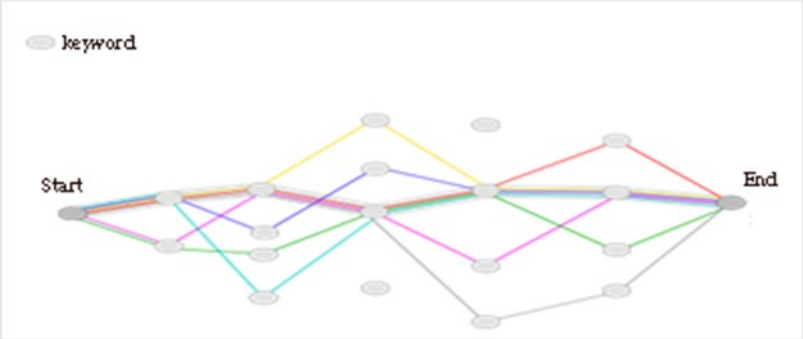


شکل ۵- مثال نمایش فضایی که کلمات کلیدی در آن قرار دارند

برای استفاده از فرمول تصور کنید که جمله زیر را داریم و کلمات کلیدی "تهران" و "مدتها" شناسایی شده است.

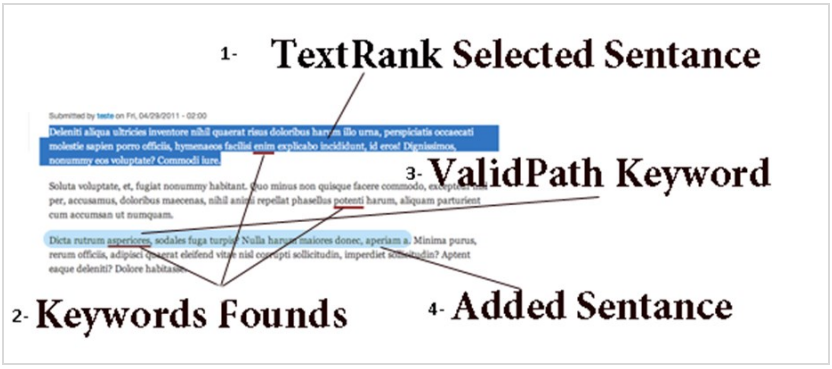
شکل۶ -یک جمله نمونه

ISSN: 2745-3673
http://www.Rcsj.ir



شکل ۷- نمایش فرضی پیدا کردن مسیرها توسط کلنی مورچه

کلمه تهران اولین کلمه کلیدی در لیست ما است و کلمه مدتها دومین فاصله بین این کلمات ۸ کلمه است. همینطور برای کلمات کلیدی دیگر این کار را تکرار میشود تا یک ماتریس تشکیل شود. حالا ماتریس را به عنوان ورودی به الگوریتم کلنی مورچه داده شده و مسیر خروجی را دریافت میگردد. تنها مسیرهای مثبت و رو به جلو را موثر ارزیابی شده و جملات به دست آمده را در صورت عدم وجود اضافه میشود. فهرست کلمات کلیدی باید فاقد ایست واژه باشد.



شکل ۸- جملات و کلمات در دو مرحله پیدا شده و افزوده میشود

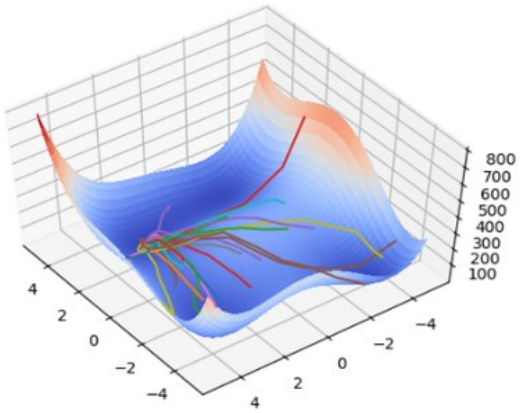
در برنامه ایجاد شده تعداد مورچه ها ۴ عدد تاثیر فرومون در جهت ۳ تاثیر فاصله گره مجاور۲ و فاکتور کاهش دهنده فرمون ۰.۰۱ و فاکتور افزایش فرمون ۲ درنظر گرفته شده است و نهایتا الگوریتم دارای فاکتور بیشترین تکرار است که ۱۰۰۰۰ در نظر گرفته شده است و مطابق شکل ۳-۴ مورد استفاده قرار گرفته است.

```
while (time < maxTime)
        {
         UpdateAnts(ants, pheromones, dists);
         UpdatePheromones(pheromones, ants, dists);

         int[] currBestTrail = AntColony.BestTrail(ants, dists);
         double currBestLength = Length(currBestTrail, dists);
         if (currBestLength < bestLength)
         {
           bestLength = currBestLength;
           bestTrail = currBestTrail;
         }
         time += 1;
        }
```

شکل ۹- استفاده از تکرار در الگوریتم کلنی مورچه

ISSN: 2745-3673
http://www.Rcsj.ir



شکل ۱۰- فضای انتزاعی متصور شده در مغز برای انتقال مفهوم توسط کلمات کلیدی و مسیرهای مرتبط. برخی جملات هدف خاصی ندارند و چیزهایی مثل تعارف هستند

ارزیابی

برای آزمایش خلاصه ساز میتوان از روشهای ماشینی و انسانی استفاده کرد. در روشهای مبتنی بر ماشین همواره فرض میشود که یک معیار و مدل خلاصه شده وجود دارد. در این جا به علت در دسترس نبودن پیکره مرجع مناسب عملا نتایج با پرسش نامه بررسی شده است. برای ارزیابی ۳ متن مرجع در نظر گرفته شده است. متن اول با موضوع اینستاگرام متن دوم با موضوع افزونگی داده و متن سوم در مورد رمز ارز میباشد. از ابتدا به انتها متون تخصصی‌تر میشوند. با توجه به در دسترس نبودن پیکره پاسخ بعلاوه اینکه این پیکره از نظرات ده نفر دانشجوی کارشناسی تشکیل شده است [16] جهت رفع نقایص و ایجاد بازخور دقیقتر از یک جامعه آماری بزرگتر با ۱۸ عضو و با میانگین تحصیلات بالاتر از کارشناسی ارشد استفاده شده است. نتایج در جدول ۱-۵ آورده شده است.

جدول ۴- امتیاز کاربران به متن

متن اصلاح شده سوم	متن اصلاح شده دوم	متن اصلاح شده اول	کلمه کلیدی سوم	کلمه کلیدی دوم	کلمه کلیدی اول	متن خلاصه سوم	متن خلاصه دوم	متن خلاصه اول	میانگین
۴	4.33	۴.۱۶	۳.۸۳	۳.۷۵	۴.۲۵	۳.۷۷	۴.۳۳	۴.۲۷	از ۱ تا ۵
۸۰	86.66	۸۳.۳۳	۷۶.۶۶	۷۵	۸۵	۷۵.۵۵	۸۶.۶۶	۸۵.۵۵	درصد

شکل ۱۱- نمودار امتیاز کاربران به متن در حالات مختلف خلاصه شده

ISSN: 2745-3673
http://www.Rcsj.ir

مطابق نمودار ۱-۵ امتیاز کاربران با استفاده از بهینه ساز تولید یک خلاصه منطقی و قابل قبول میباشد. باید توجه داشت که متن سوم یک متن تخصصی در حوضه رمز ارز بوده است و برای اکثر جامعه آماری ناآشنا محسوب میشده است. علاوه بر این بهینه ساز تاثیر منفی محسوس روی کاهش امتیاز سایر متنها نداشته است.

نتیجه گیری

تلاشهای پردازش زبان بر نظرات مختلفی استوار است از جمله مفاهیم اولیه می‌توان به دستیابی، بازنمایی پایه مفهومی اشاره کرد که زیربنای همه زبان های طبیعی به شمار می‌روند. این واقعیت ساده که انسان ممکن است هر زبان طبیعی را برای مدتی در آن غوطه ور شود را بفهمند.و اینکه بتواند از آن زبان به هر زبان طبیعی دیگری که به خوبی با آن آشنایی دارند ترجمه کند، نشان می دهد که چنین مبنای مفهومی دارای واقعیت روانی است. افرادی که به بسیاری از زبان ها مسلط هستند می توانند آزادانه از یکی به دیگری عبور کنند، گاهی اوقات حتی بدون اینکه آشکارا بدانند در یک لحظه به چه زبانی صحبت می کنند. کاری که آن ها انجام می دهند، فراخوانی بسته ای از قوانین نقشه برداری برای یک زبان معین از پایه مفهومی است. مبنای مفهومی، محتوای اندیشه ای را دارد که بیان می شود. سپس این محتوای مفهومی از طریق قوانین تحقق به واحدهای زبانی ترسیم می شود[30] .

اما امروز ثابت شده است که رابط بین زبان و شناخت را می توان در سطوح مختلف تشخیص داد [31]. نشان داده شده است که زبان ممکن است بر روی اندیشه تاثیر بگذارد. همینطور مشخص شده است، در حوزه های ادراکی/شناختی خاصی، مرزهای مقوله های مفهومی به شدت با مرزهای معنایی اصطلاحات زبانی مربوطه همبستگی دارد. تعداد فزاینده‌ای از کارهای تجربی با برجسته نمودن تفاوت‌های داده های بین زبانی در حوزه‌های مختلف، از جمله رنگ [32]، اعداد [33]، فضا-زمان [34] و حالات ذهنی [35] از این دیدگاه پشتیبانی می‌کنند.

با این‌حال، علاوه بر سوگیری های معنایی، واضح است که استفاده مکرر از ساختارهای نحوی خاص ممکن است چالش های شناختی خاصی را برای گویندگان تحمیل کند یا عادات پردازش خاصی را تقویت کند، که در دراز مدت ممکن است روش های خاص پردازش اطلاعات را فراتر از حوزه زبانی افزایش دهد [36].

در این مقاله، با توجه به مطالعات و تحقیقات پیشین روشی برای استخراج کلمات کلیدی متن و خلاصه سازی ارائه شد. اساس این روش بر این موضوع است که انسانها سعی در انتقال مفاهیم در کوتاه ترین زمان را دارند. اما تأثیر اصلی زبان بر اندیشه احتمالاً ناشی از عادت کردن به استراتژی‌هایی است که برای درک، تفسیر و یادآوری دنیایی که ما را احاطه کرده است میباشد.

ISSN: 2745-3673
http://www.Rcsj.ir

منابع و مراجع

[1] M. V. V. V. M. S. I. Ashmarina, Digital age: chances, challenges and future, 2020.
[2] M. Rosemann, "Structuring in the Digital Age," The Art of Structuring, 2019.
[3] H. Dalianis, "A Text Summarizer for Swedish," NADA, KTH, Stockholm, 2000.
[4] M. Hassel and N. Mazdak, "FarsiSum - A Persian Text Summarizer," Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, p. 82–84, 2004.
[5] Azadeh Zamanifar, Behrouz Minaei and Mohsen Sharifi, "A New Hybrid Farsi Text Summarization Technique Based on Term Co-Occurrence and Conceptual Property of the Text," Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2008.
[6] T. A. a. M. E. J. Mehrnoush Shamsfard, "Persian Document Summarization by Parsumist," World Applied Sciences Journal 7 (Special Issue of Computer & IT): 1, 2009.
[7] A. Zamanifar and O. Kashefi, "AZOM: A Persian Structured Text Summarizer," Natural Language Processing and Information Systems, p. 234–237, 2011.
[8] Fatemeh Shafiee and Mehrnoush Shamsfard, "Similarity versus relatedness: A novel approach in extractive Persian document summarisation," Journal of Information Science, 2017.
[9] Tayyebeh Hosseinikhah, Abbas Ahmadi and Azadeh Mohebi, "A new Persian Text Summarization Approach based on Natural Language Processing and Graph Similarity," Iranian Journal of Information Processing and Management, 2018.
[10] F. Kiyoumarsi and F. Esfahani, "Optimizing Persian Text Summarization Based on Fuzzy Logic Approach," International Conference on Intelligent Building and Management, 2011.
[11] M. Tofighy, O. Kashefi and H. Javadi, "Persian Text Summarization Using Fractal Theory," in Communications in Computer and Information Science ·, 2011 .
[12] M. Bazghandi, G. Tadayon, T. Jahan and M. Vafaei, "Extractive Summarization of Farsi Documents Based on PSO Clustering," IJCSI International Journal of Computer Science, 2012 .
[13] Seyyed Mohsen Tofighy, Ram Gopal Raj and Hamid Haj Seyyed Javad, "AHP Techniques for Persian Text Summarization," Malaysian Journal of Computer Science, 2013.
[14] Asef Pour masoomi, Mohsen Kahani, Seyyed Ahmad Toosi and Ahmad Estiri, "Ijaz: An Operational system for single-document summarization of Persian news texts," Signal and Data Processing, 2014 .
[15] T. Strutz, Data Fitting and Uncertainty A practical introduction to weighted least squares and beyond, Leipzig, Germany, 2010 .
[16] B. B. Moghaddas, M. Kahani, S. A. Toosi, AsefPourmasoumi and A. Estiri, "Pasokh: A Standard Corpus for the Evaluation of Persian Text Summarizers," 3rd International Conference on Computer and Knowledge Engineering, 2013 .
[17] Saeed Farzi and Sahar Kianian, "Katibeh: A Persian news summarizer using the novel semi-supervised approach," Digital Scholarship in the Humanities, 2018 .
[18] Mohammad Fakhredanesh, Mohammad Ebrahim Khademi and Seyed Mojtaba Hoseini, "Farsi Conceptual Text Summarizer: A New Model in Continuous Vector Space ."
[19] H. M. H. I. A. A. M. Al-Zahrani, "PSO-Based Feature Selection for Arabic Text Summarization," Computer Science, 2015 .
[20] Mohammed Binwahlan, Naomie Salim and Ladda Suanmali, "Swarm Based Text Summarization," International Association of Computer Science and Information Technology, 2009 .
[21] S. Miles, L. Yao, W. Meng, C. M. Black and Z. B. Miled, "Topic Extraction from A Cancer Health Forum," IEEE 9th International Conference on Healthcare Informatics (ICHI), 2021 .
[22] Samuel Miles, Lixia Yao, Weilin Meng, Christopher M. Black and Zina Ben Miled, "Comparing PSO-based clustering over contextual vector embeddings to modern topic modeling," Information Processing & Management, vol. 59, no. 3, 2022 .
[23] Shrabanti Mandal, Girish Kumar Singh and Anita Pal, "Single document text summarization technique using optimal combination of cuckoo search algorithm, sentence scoring and sentiment score," International Journal of Information Technology, vol. 13, p. 1805–1813, 2021 .
[24] R. Z. Al-Abdallah and A. T. Al-Taani, "Arabic Text Summarization using Firefly Algorithm," IEEE, 2019 .

ISSN: 2745-3673
http://www.Rcsj.ir

[25] Kaleab Getaneh Tefrie and Kyung-Ah Sohn, "Autonomous Text Summarization Using Collective Intelligence Based on Nature-Inspired Algorithm," International Conference on Mobile and Wireless Technology, p. 455–464, 2017 .

[26] Jesus M. Sanchez-Gomez, Miguel A. Vega-Rodríguez and Carlos J. Pérez, "Extractive Multi-Document Text Summarization Using a Multi-Objective Artificial Bee Colony Optimization Approach," Knowledge-Based Systems, vol. 159, 2017 .

[27] Richa Sharma, Sudha Morwal and Basant Agarwal, "Named entity recognition using neural language model and CRF for Hindi language," Computer Speech & Language, vol. 74, 2020 .

[28] Ahmad Alhasan and Ahmad T. Al-Taani, "POS Tagging for Arabic Text Using Bee Colony Algorithm," Procedia Computer Science, vol. 142, 2018 .

[29] Ling Zhao, Ailian Zhang, Ying Liu and Hao Fei, "Encoding multi-granularity structural information for joint Chinese word segmentation and POS tagging," Pattern Recognition Letters, vol. 138, pp. 163-169, 2020 .

[30] Roger C. Schank, "Conceptual dependency: A theory of natural language understanding," Cognitive Psychology, pp. 552-631, 1972 .

[31] E. &. A. F. T. W. Hunt, "The Whorfian hypothesis: A cognitive psychology perspective.," Psychological Review, p. 377–389, 1991 .

[32] Aubrey L. Gilbert, Terry Regier, Paul Kay and Richard B. Ivry, "Whorf hypothesis is supported in the right visual field but not the left," Biological Sciences, pp. 489-494, 2005 .

[33] Rochel Gelman and C. R. Gallistel, "Language and the Origin of Numerical Concepts," Science, pp. 441-443, 2004 .

[34] Rafael E. Núñez and Eve Sweetser, "With the Future Behind Them: Convergent Evidence From Aymara Language and Gesture in the Crosslinguistic Comparison of Spatial Construals of Time," Cognitive Science Society, Inc, 2005 .

[35] Jennie E. Pyers and Ann Senghas, "Language Promotes False-Belief Understanding: Evidence From Learners of a New Sign Language," Psychological Science, 2009 .

[36] Simpson-Finch H, Yohe Moore ES, Brandt B, Poepsel T, Heinzman A and Dempsey S, "PMU61 - Linguistic and Cultural Considerations When Implementing A Global ‘Bring your Own Device’ (BYOD) Study," Value in Health, vol. 21, pp. 590-591, 2018 .

[37] Ángel Hernández-Castañeda, René Arnulfo García-Hernández, Yulia Ledeneva and Christian Eduardo Millán-Hernández, "Language-independent extractive automatic text summarization based on automatic keyword extraction," Computer Speech & Language, vol. 71, 2022 .

[38] Abdhul Ahadh, Govind Vallabhasseri Binish and Rajagopalan Srinivasan, "Text mining of accident reports using semi-supervised keyword extraction and topic modeling," Process Safety and Environmental Protection, vol. 155, pp. 455-465, 2021.