خلاصه‌ساز متون روایی مبتنی بر جنبه‌های شناختی ذهن انسان

سیّده ساره صادقی و بهرام وزیرنژاد

گروه زبان‌شناسی رایانشی، مرکز زبان‌ها و زبان‌شناسی، دانشگاه صنعتی شریف، تهران، ایران

چکیده

این پژوهش به طراحی و ایجاد یک سامانهٔ خلاصه‌ساز متن که مبتنی بر نظریات شناختی است، می‌پردازد. نظریهٔ مدل موقعیتی، مهم‌ترین نظریه در زمینه عملکرد ذهن در درک متن شناخته می‌شود که برای توضیح فرآیند درک متون روایی کاربرد دارد. در مقایسه با دیگر روش‌ها که به‌طور معمول رویکردی آماری دارند، این روش از این نظر که سامانه‌ای مبتنی بر عملکرد ذهن انسان است، روش نوینی محسوب می‌شود؛ به‌علاوه توانسته یکی از نظریات شناختی معتبر را در قالب یک سامانهٔ خودکار خلاصه‌سازی به بوته آزمایش گذارد. نظریهٔ شناختی مدل موقعیتی، بیان می‌کند که به هنگام خوانش متن، انسان به پنج ویژگی در سطح جمله، شامل تغییر زمان و مکان، روابط علّی، میزان ارتباط با موضوع و شخصیت‌های اصلی توجه و چارچوبی ذهنی بر حسب آن ایجاد می‌کند تا جملات مهم متن شناسایی شوند. نتایج به‌دست‌آمده برای این روش دقتی رضایت‌بخش و قابل مقایسه با روش‌های روز آماری را به‌دست داده که از این جهت که مبتنی بر یک نظریهٔ شناختی درک متن است، حائز اهمیت می‌باشد.

واژگان کلیدی: خلاصه‌سازی خودکار استخراجی، علوم شناختی، نظریه مدل موقعیتی[1]، متون روایی.

۱- مقدّمه

گسترش روزافزون منابع فضای مجازی و نیازی که هر روز برای صرفه‌جویی در وقت و هزینه افراد احساس می‌شود، پژوهش‌های بسیاری را در زمینه خلاصه‌سازی متون موجب شده است. بیشتر قریب به اتفاق طرح‌های ارائه شده در این حوزه، رویکردی آماری را برگزیده و کم‌تر به روش‌های شناختی که به عمل‌کرد ذهن و مغز انسان توجه دارد، پرداخته می‌شود.

در سال‌های اخیر، همکاری دانشمندان حوزه‌های زبان‌شناسی، علوم شناختی، فلسفهٔ ذهن، عصب‌شناسی و هوش مصنوعی سبب شده است تا گام‌های مهمی در حوزه‌های بین رشته‌ای نظیر زبان‌شناسی شناختی برداشته شود. همچنین در سال‌های اخیر در سطح نخست پژوهش‌های جاری در دنیا در حوزهٔ پردازش زبان طبیعی به عملکرد ذهن در درک متن توجه شده است؛ اما همچنان راه بسیاری برای رسیدن به وضعیت مطلوب در این علوم وجود دارد. در ایران با وجود پیشرفت‌هایی که در سایر علوم اتفاق افتاده، متأسفانه فعالیت بسیار زیادی در حوزهٔ علوم شناختی و تلفیق و استفاده از آن در پردازش زبان طبیعی صورت نگرفته است و همین سبب شد تا در پی طراحی سامانه‌ای خلاصه‌ساز، مبتنی بر عمل‌کرد شناختی ذهن انسان باشیم که بی‌تردید می‌تواند به گامی در تحول بینش مهندسان پردازش زبان طبیعی در حل مسائل این حوزه منجر گردد و همچنین می‌تواند به درک بهتر ما از عملکرد ذهن در پردازش زبان کمک نماید.

در این مقاله ابتدا در بخش ۲ مروری بر کارهای مرتبط انجام شده خواهیم داشت؛ سپس در بخش ۳ نظریات شناختی مرتبط با موضوع، به‌خصوص نظریاتی که مبنای طراحی سامانه خلاصه ساز مورد نظر هستند ارائه می‌شود. در بخش ۴ به معماری سامانه طراحی‌شده با ذکر جزییات و زیرسامانه‌ها پرداخته می‌شود؛ سپس در بخش ۵ به ارزیابی

[1] situational model theory

سامانه و خروجی آن پرداخته می‌شود و با خلاصه‌ساز مطرح دیگری مقایسه خواهد شد. بخش ۶ نیز به بحث و نتیجه‌گیری اختصاص دارد.

۲- مروری بر کارهای گذشته

نخستین سامانه‌های خلاصه‌سازی توسط *جان لوهن* (لوهن، ۱۹۵۸) و *ادموندسون* (ادموندسون، ۱۹۶۸) با توجه به مشخصات متن ساخته شد. در این کارها به ویژگی‌هایی از قبیل بسامد کلمات، مکان جملات، میزان شباهت با عنوان، طول جمله، وجود اسامی خاص در جمله، وجود کلماتی با حروف بزرگ، وجود کلماتی با قلم متفاوت از قلم بقیه کلمات، وجود اعداد در جمله و محاسبه میزان شباهت هر جمله با دیگر جملات توجه شده بود. به این ترتیب با نمره‌دهی به جملات برحسب ویژگی‌ها، جملات مهم برای سامانه شناسایی می‌شد. بعد از آن زمان، از دهۀ ۱۹۹۰ با گسترش اینترنت و توجه روزافزون به منابع الکترونیکی کارهای بهتری در زمینه خلاصه‌سازی انجام شد. از سال ۱۹۹۰ به بعد استفاده از زنجیرۀ واژگانی با بهره‌گیری از هستان‌شناسی‌ها نظیر وردنت[1] (میلر، ۱۹۹۵ ص۴۱-۳۹) مورد توجه قرار گرفت. در این روش‌ها سعی بر آن بود تا با تشکیل زنجیرۀ واژگانی و روابط معنایی که بین جملات و یا بین کلمات وجود داشت، بخش‌های مهم متن شناسایی شوند؛ سپس با روی کار آمدن روش‌های ماشین‌یادگیری، خیلی زود این روش‌ها برای طراحی سامانۀ خلاصه‌ساز استفاده شدند. در این کارها، مدل نایو بیز (اریس، ۲۰۱۳) و مدل مخفی مارکوف (فیلاتوا، ۲۰۰۴) و تحلیل معنایی پنهان (دیروستر، ۱۹۹۰) از پرطرفدارترین کارها بودند. در کار انجام‌شده برای مدل نایو بیز، از خود متن به‌عنوان پیکرۀ آموزشی استفاده می‌کند و پس از پیش‌پردازش متون با استفاده از روش‌های اندازه‌گیری شباهت نظیر شباهت کسینوسی، به خوشه‌بندی جملات مشابه می‌پردازد. بدین ترتیب جملات در دسته‌های موضوعی یکسان قرار می‌گیرند. در مرحلۀ بعدی، با استفاده از الگوریتم دسته‌بندی‌کنندۀ نایو بیز هر خوشه را به‌عنوان یک طبقه درنظر گرفته و ویژگی‌های هر خوشه را شناسایی می‌کند؛ سپس به قصد انتخاب جملاتی که بتواند بیش‌ترین موضوعات را نشان دهد، برحسب ویژگی‌های به‌دست آمده به جملات نمره داده می‌شود و جملات با نمرات بالا انتخاب می‌شوند.

ویژگی مدل مخفی مارکوف آن است که موضوع‌محور و مبتنی بر حوزۀ خاص متون است. بدین ترتیب حالت‌های موجود در ساختار مدل و نیز معماری مدل وابسته به حوزۀ خاص متن است. برای مثال در متون مرتبط با موضوع زلزله، یک ساختار مارکوف با حالت‌های "مکان وقوع"، "زمان وقوع"، "شدت زلزله" و "میزان خرابی" ساخته شده و سند ورودی روی الگوی مارکوف پذیرش می‌شود به‌نحوی که هر جمله به یک حالت تخصیص می‌یابد، درنهایت از هر حالت مدل، تعدادی از جملات که بیش‌ترین تطبیق را با آن حالت دارند، برگزیده می‌شوند. مدل مخفی مارکوف برای متون داستانی، به‌خوبی عمل کرده و برای انتخاب محتوا و ترتیبِ آمدن جملات قابلیت خوبی دارد.

در سال‌های ۲۰۰۴ و ۲۰۰۵ روش‌های مبتنی بر گراف، مورد توجه قرار گرفت (مانی، ۱۹۹۷). در این روش‌ها جملات تنها به محتوای خاصی تعلق نداشت، بلکه برای مثال می‌توانست دارای محتوایی شامل هم زمان وقوع و هم مکان وقوع و نظایر آن باشد. در مدل مبتنی بر گراف، تعدادی گره موجود است که نمایان‌گر جملات یا بندها هستند. در این مدل‌ها ارتباط بین گره‌ها برحسب میزان شباهت و ارتباطشان با یکدیگر، برحسب ویژگی‌های تعیین شده مشخص می‌شد.

برخی سامانه‌های خلاصه‌سازی، به‌طور تخصصی برای حوزه‌های خاص هم‌چون ویکی‌پدیا (ساوپر، ۲۰۰۹)، زندگی‌نامه (اسچیف‌من، ۲۰۰۱) و صفحات وب (تورپین، ۲۰۰۷) طراحی شده‌اند. در این سامانه‌ها به ویژگی‌های نگارشی و ساختاری که برای نوشتن متن استفاده می‌شد، توجه شده است.

در سال‌های اخیر بیش‌ترین توجه بر آن است تا از جملات حشو در متون خلاصه‌شده در سامانه جلوگیری شود؛ به همین دلیل از روش‌هایی مانند بیش‌ترین ارتباط مرزی استفاده شده که مبتنی بر گراف است (لین، ۲۰۱۰)؛ علت آن نیز نمایش طبیعی روابط و تعامل واحدهای متنی است؛ سپس در فرمولی که برای تعیین میزان شباهت جملات استفاده می‌شود، مقدار امتیاز منفی در نظر گرفته می‌شود تا از جملاتی که سامانه، مشابه شناسایی کرده است، کاسته شود.

در طول این نیم‌قرنی که بر روی سامانه‌های خلاصه‌ساز کار شده است، هنوز خلاصه‌سازی متون مسأله‌ای حل‌نشده و به‌عنوان یکی از پرچالش‌ترین مسائل در حوزۀ

[1] Wordnet

پردازش زبان طبیعی مطرح و راه زیادی تا رسیدن به وضعیت رضایت‌بخش باقی‌مانده است.

۳- مروری بر نظریات شناختی

نظریهٔ مدل موقعیتی، یک نظریهٔ شناختی در مورد عملکرد ذهن است که چارچوب اطلاعاتی از وضعیت امور یک جهان محدودشده در ذهن انسان را به‌دست می‌دهد (وان دیجک، ۱۹۸۳). بر طبق زیربخش درک متن از این نظریه، به هنگام خوانش متن توسط خواننده، اطلاعاتی از قبیل زمان، مکان، توالی علّی وقایع، میزان ارتباط جمله با موضوع و شخصیت‌های اصلی از متن کسب می‌شود. این ویژگی‌ها اجزای سازندهٔ نظریه مدل موقعیتی هستند. در حین مطالعهٔ پیوسته، اطلاعات جدید موجود متن در قالب چارچوب قبلی به تصویر ذهنی اضافه می‌شود. مدل نمایه‌ساز رخدادها[1] که گونه‌ای از مدل موقعیتی است، روی پنج عنصر مدل موقعیتی تمرکز کرده و مخصوص متون روایی یا داستانی است. روش انتخابی در این پژوهش نیز بر پایهٔ همین مدل می‌باشد. در این بخش به ویژگی‌های پنج‌گانهٔ مورد توجه در این مدل می‌پردازیم که برای هر جمله در نظر گرفته می‌شوند.

۳-۱- ویژگی زمان

در متون روایی، حوادث در سیری با توالی زمانی بیان می‌شوند؛ بدین معنی که حوادث در سیری از زمان به وقوع پیوسته و همان پایه‌ای از تصویر ذهنی را می‌سازد. اطلاعات جدید و یا هرگونه تغییری که در این سیر زمانی به‌وجود می‌آید، سبب به‌روزشدن تصویر ذهنی می‌شود.

۳-۲- ویژگی مکان

عنصر مکان در تصویر ذهنی انعکاس و بازسازی مکان رخدادها در متن است (هارت، ۱۹۸۳) و بدین ترتیب ذهن انسان اطلاعات مکانی را در درون خود نگه می‌دارد. فرایندهای ذهنی- مکانی را می‌توان با شناسایی مکان‌ها و انتقالاتی که صورت می‌گیرد، توصیف کرد.

۳-۳- ویژگی علّیت یا توالی علّی رخدادها

هر ساختمان علیت، شامل دو بخش است؛ یک علّت و دیگری معلول (گیرجو، ۲۰۰۳). علت پدیده‌ای مجزا از معلول است و بر معلول نیرویی را به شکل فیزیکی و یا غیر فیزیکی وارد می‌کند. معلول پدیده‌ای است که وضعیتش به علت نیروی واردشده تغییر می‌کند. جملاتی که شرطی و استدلالی هستند و یا رابطه علت-معلولی را به‌صورت غیر صریح بیان می‌کنند، جزو دستهٔ جملات علّی قرار می‌گیرند (بلانکو، ۲۰۰۸). درصورتی‌که دو عنصر علّت[2] و معلول[3] در جمله باشند، رابطهٔ علّی به‌صورت صریح و آشکار است؛ اما اگر یکی از آن دو نباشند، رابطه به‌صورت تلویحی است. به این مثال توجه کنید: "جواد از مهمان‌سرا اخراج شد به‌دلیل آن‌که یکی از مشتریان از وی شکایت کرد." همان‌گونه که مشاهده می‌شود هر دو عنصر علّت و معلول در جمله حضور دارند. مدل‌های نوین علّیت مفهوم علّیت را در قالبی توسعه‌یافته از مفاهیم که علاوه‌بر موارد بالا شامل مفاهیمی همچون اجازه‌دادن[4]، کمک‌کردن[5]، جلوگیری‌کردن[6] و مانع‌شدن[7] است گسترش می‌دهند (ولف، ۲۰۰۳).

روشن است که استخراج روابط صریح و آشکار، بسیار ساده‌تر از روابط علّی تلویحی است. الگوهای علّی آشکار شامل کلماتی همچون تأثیرگذاشتن، منجرشدن و غیره است. روابط تلویحی علّی پیچیده‌تر از روابط علّی آشکار هستند و نیاز به تحلیل معنایی و دانش جهان بیرون دارند (گیرجو، ۲۰۰۳). برای روابط علّی تلویحی، نشانه‌ای در متن وجود ندارد، بنابراین کشف آنها توسط ماشین‌ها دشوار است و توسط انسان‌ها با استنتاج و استنباط ذهنی مشخص می‌شوند. اکثر کارهای انجام‌شده در کشف روابط علّی به‌صورت خودکار بر کشف روابط صریح تمرکز کرده‌اند. خوشبختانه نتایج آزمایش‌های انجام‌شده توسط ما و دیگران (گیرجو، ۲۰۰۳) نشان داده است که استفاده از همین بخش صریح از روابط علی می‌تواند نتایج رضایت‌بخشی در ایجاد خلاصه خودکار از متن داشته باشد.

۳-۴- ویژگی شخصیت اصلی جمله

شخصیت اصلی هر جمله به معنی فاعل جمله و همچنین شخصیت اصلی جمله به معنای انجام‌دهنده کار است؛ و موضوع جمله بر حول شخصیت او می‌چرخد (زوان، ۱۹۹۵). در این کار جملات براساس شخصیت‌های اصلی خوشه‌بندی می‌شوند.

[1] event indexing model

[2] affecter
[3] patient
[4] letting
[5] helping
[6] preventing
[7] hindering

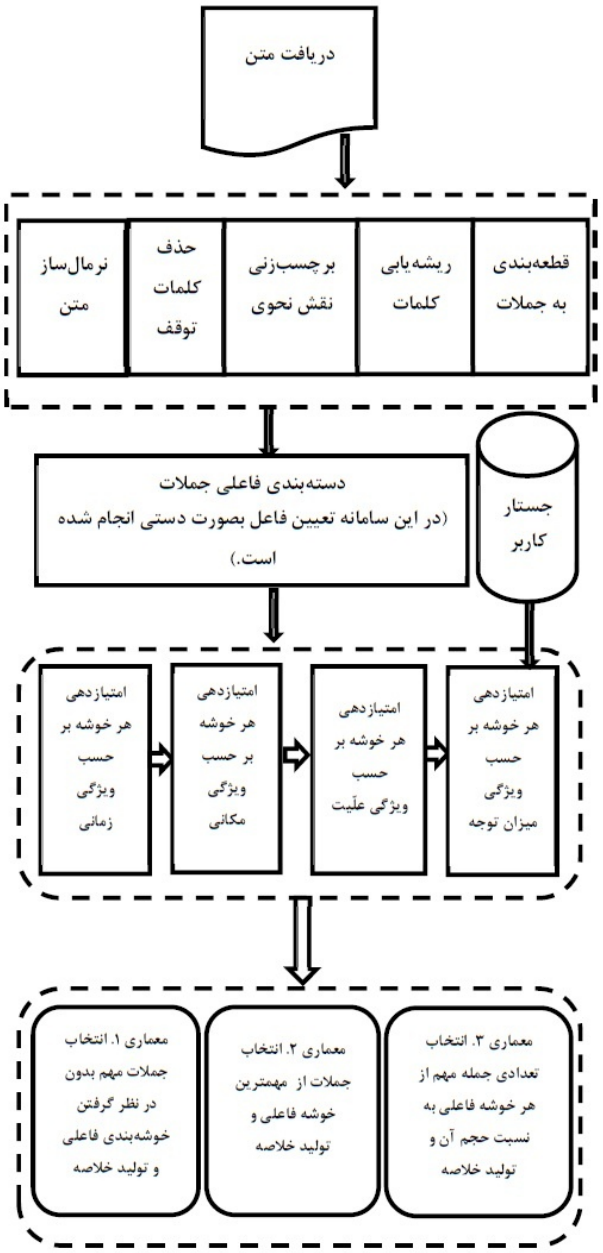
۳-۵- ویژگی میزان توجه

یکی از عناصر سازندهٔ تصویر ذهنی ما از متن که در سطح جمله مفهوم می‌یابد، میزان ارتباط جمله با هدف اصلی نگارنده است. خواننده نیز در هنگام خوانش متن روی جملاتی که حائز چنین ویژگی باشد به‌صورت خودآگاه و یا ناخودآگاه بیشتر متمرکز شده و به پردازش بیشتر آنها می‌پردازد (کیفه، ۱۹۹۳). در حالت ناخودآگاه این توجه غیر ارادی بوده و خواننده متوجه عملی که انجام می‌دهد نیست. گاه نیز با طرح پرسشی برای خواننده این نوع توجه به‌صورت آگاهانه و هوشیارانه و به‌صورت ارادی است.

۴- روش کار و معماری سامانه

روش ارائه‌شده برای این سامانه استفاده از مدل نمایه‌ساز رخدادهاست. همان‌گونه که در شکل (۴-۱) مشخص شده است، در سامانهٔ ارائه‌شده، پس از مراحل پیش‌پردازش متن، چهار ویژگی مدل نمایه‌ساز رخدادها برای هر جمله استخراج می‌شود؛ سپس برحسب میزان اهمیت ویژگی‌های تعریف‌شده از مجموع ویژگی‌های استخراج‌شده برای هر جمله نمره‌ای حاصل می‌شود. ارزش‌های اختصاص‌داده‌شده به این ویژگی‌ها بدین صورت مشخص شده‌اند که چون در متن، روابط بین جملات و کلمات، دارای ارزش یکسانی نیست پس به روابط علی و میزان توجه ارزش بیش‌تری اختصاص داده می‌شود. این ارزش‌ها بدین گونه هستند: تغییر مکانی و زمانی ارزشی برابر با "۱"، داشتن روابط علی ارزشی برابر با "۲" و برای میزان شباهت با جستار[1] از مقدار به‌دست‌آمده از شباهت کسینوسی استفاده می‌شود؛ منظور از جستار، عبارت یا مجموعه‌ای از کلمات است که توسط کاربر وارد می‌شود تا خلاصه‌ای مرتبط با آن جستار تولید شود، در مورد متونی که وجوه موضوعات مختلفی دارند، جستار امکان ایجاد خلاصه‌ای مرتبط با یکی از موضوعات را به کاربر می‌دهد؛ سپس برحسب یکی از سه روش پیشنهادی، جملات مهم انتخاب می‌شوند و در خلاصه به‌ترتیب حضور در متن ورودی ظاهر می‌شوند.

همان‌گونه که مشخص است در مرحله انتخاب جملات مهم نیز سه روش برحسب میزان تأثیر فاعل پیشنهاد شده است. در روش نخست تنها جملات مهم بدون توجه به فاعل انتخاب می‌شوند؛ در روش دوم با توجه به میزان نمرات هر خوشه فاعلی مهم‌ترین خوشه انتخاب شده است و تمام جملات موجود در آن با توجه به حجم مطلوب به‌عنوان خلاصه در نظر گرفته می‌شوند در روش سوم از هر خوشه فاعلی بر حسب حجمی که در متن اشغال کرده است جملات مهم انتخاب می‌شوند..



(شکل ۴-۱): معماری کلی برای سامانه پیشنهادی

[1] Query

۴-۱- پیش‌پردازش

در مرحله پیش‌پرداز به پردازش‌هایی هم‌چون نرمال‌سازی، ریشه‌یابی، برچسب‌زنی نحوی و حذف کلمات توقف پرداخته شده است.

۴-۱-۱- نرمال‌سازی متن

از جمله چالش‌هایی که در زبان فارسی برای سامانه‌های خلاصه‌سازی وجود دارد، مشکلات نگارشی زبان فارسی است. برای نمونه می‌توان به استفادهٔ نابه‌جای فاصله به‌جای نیم‌فاصله اشاره کرد. توجه به استفاده از علایم نگارشی فارسی و استفاده از حروف غیر عربی می‌تواند ایرادهای نگارشی متن را کم کند. از دیگر موارد می‌توان وجود کلمات چندین املایی را ذکر کرد. هم‌چنین شناسایی کلمات مرکب و درنظرگرفتن آنها به‌عنوان یک کلمه از موارد دیگر است. در این پژوهش از نرم‌افزار طراحی‌شده در آزمایشگاه پردازش گفتار و زبان، دانشگاه صنعتی شریف استفاده شد. این نرم‌افزار براساس آیین نگارش مصوب فرهنگستان زبان و ادب فارسی تهیه شده است و قادر به تنظیم فواصل و نیم‌فاصله‌ها با توجه به این قوانین است. همچنین قادر به تبدیل نویسه‌های عربی به فارسی و تصحیح غلط‌های املایی و تایپی است. جزییات کامل و نحوه عملکرد این سامانه در (سلطان‌زاده، ۱۳۹۲) گزارش شده است.

۴-۱-۲- برچسب‌زنی نقش نحوی کلمات

برچسب‌زنی نحوی کلمات به دو منظور انجام گرفته است. نخست آنکه در هنگام ریشه‌یابی با توجه به مقوله نحوی کلمه، ریشه‌یابی صورت می‌گیرد. به‌عنوان مثال، قواعد تصریفی فعل برای همان دستهٔ نحوی و قواعد تصریفی اسم برای همان دسته در نظر گرفته می‌شوند. دومین دلیل نیاز به برچسب نحوی، استخراج ویژگی زمان هر جمله است که با توجه به برچسب‌داربودن کلمات، تنها به افعال توجه می‌شود و با توجه به ریشهٔ آنها زمان جمله مشخص می‌شود. این برچسب‌زن مطابق با دادهٔ آموزشی (بی‌جن‌خان، ۲۰۱۱) آموزش داده شده و کلمات را در دسته‌های مشخص‌شدهٔ اسم، فعل، قید، حرف اضافه، صفت، حرف ربط، کلمات شرطی، نشانه مفعولی (را)، ضمایر، وابسته‌های اسمی و علایم نگارشی تقسیم‌بندی می‌کند.

۴-۱-۳- ریشه‌یابی کلمات

مهم‌ترین نیاز به ریشه‌یابی کلمه در این خلاصه‌ساز، تعیین ریشهٔ افعال به‌منظور تعیین زمان جمله است. هم‌چنین در امتیازدهی به جملات از نظر میزان شباهت با جستار قبل از محاسبه شباهت کسینوسی تمامی اسامی و افعال ریشه‌یابی شده و از ریشه‌ها در ایجاد بردار هر جمله استفاده می‌شود. اگرچه هر وند تصریفی برای طبقهٔ خاصی از کلمات استفاده می‌شود و در هر مکانی نمی‌تواند قرار بگیرد و جایگاه مشخصی دارد، با این حال ریشه‌یابی خودکار کلمات کاری چالش برانگیز است؛ با این وجود در این پژوهش سعی بر آن شد تا ریشه‌یاب، در زمینهٔ ریشه‌یابی افعال، عملکرد صددرصدی داشته باشد.

۴-۱-۴- حذف کلمات توقف متن

کلمات توقف حدود پنجاه کلمه، شامل کلمات دستوری نظیر حروف ربط، حروف اضافه و غیره که فاقد اطلاعات معنایی و محتوایی مرتبط با متن هستند. ویژگی این کلمات این است که در هر نوع متن با هر موضوعی به‌شکل یک‌سانی دیده می‌شوند. در این سامانه پیش از پردازش اصلی متن، کلمات توقف از متن ورودی حذف می‌شوند. فهرستی از کلمات توقف موجود در پیکرهٔ همشهری برای این کار استفاده شد (آل‌احمد، ۲۰۱۱).

۴-۱-۵- قطعه‌بندی به جملات

برای تعیین ویژگی‌های گفته‌شده برای هر جمله، متن ورودی به جملات تشکیل‌دهنده‌اش تقسیم می‌شود. قطعه‌بندی جملات در این سامانه با استفاده از علائم سجاوندی و درنظرگرفتن موارد استثنا، نظیر استفاده از نقطه در کوته‌نوشت‌ها تهیه شده است. این قطعه‌بند در عین سادگی از عملکرد بسیار خوبی برخوردار است.

۴-۲- دسته‌بندی فاعلی جملات

شخصیت‌های اصلی به‌دلیل آنکه سبب وقوع حوادث در سیر زمانی و مکانی در متون روایی می‌شوند، نقش مهمی در متن و خلاصه متن ایفا می‌کنند. به همین جهت، در این سامانه توجه ویژه‌ای به فاعل جملات شده است. برای تعیین فاعل، نیازمند ابزاری برای ایجاد درخت نحوی بودیم؛ همچنین تعیین مرجع ضمایر، که در بسیاری از اوقات به‌جای عبارات اسمی می‌نشینند، به‌طور خودکار چالش بزرگی در زبان

فارسی است؛ چون مرجع ضمیر در بسیاری اوقات دارای ابهام است. نتیجهٔ بررسی نگارندگان در این قسمت این است که ابزاری قابل اتکا در این بخش در زبان فارسی در حال حاضر موجود نیست که بتواند در متون روایی عملکرد مناسبی داشته باشد؛ وانگهی ایجاد چنین سامانه‌ای هم نیازمند کاری پرحجم و پژوهشی بلند مدت است؛ لذا از آن جهت که هدف اصلی این پژوهش بررسی، مطالعه و درصورت امکان ارائه و ایجاد چارچوبی مبتنی بر مدل‌های شناختی برای یک سامانه خلاصه‌ساز خودکار بود، تصمیم گرفته شد تا به سامانه خلاصه‌ساز در این بخش کمک شود؛ بنابراین با توجه به ذات زبان فارسی که ضمیر در آن زیاد استفاده می‌شود و این‌که ضمایر در زبان فارسی از نظر جنسیت نشان‌دار نیستند و ابهام در مرجع ضمایر در زبان فارسی و نیاز به شناسایی مراجع آن‌ها و فقدان منابع مناسب برای شناسایی خودکار روابط معنایی کلمات با یکدیگر، اطلاعات فاعلی به‌صورت دستی به سامانه ارائه شد؛ بدین صورت که فاعل هر جمله به‌صورت دستی در متن ورودی مشخص و بعد در اختیار سامانه قرار می‌گیرد. جملات با توجه به شمارهٔ فاعل آنها به دسته‌های فاعلی خوشه‌بندی می‌شوند تا بعداً طبق معماری نمایش‌داده‌شده در شکل (۴-۱) از هر خوشه تعدادی جمله برای ایجاد خلاصه استفاده شود.

۴-۳- استخراج ویژگی‌های هر جمله مبتنی بر مدل موقعیتی

پس از مراحل پیش‌پردازشی که متن را برای پردازش اصلی آماده می‌کنند و قراردادن جملات در دسته‌های فاعلی، از دیگر عناصر سازندهٔ موجود در چارچوب نظری مدل موقعیتی به‌عنوان ویژگی‌های امتیازدهنده به جملات استفاده می‌شود تا براساس امتیازات کسب‌شده جملات دارای امتیازهای بیشتر برای ایجاد خلاصه به‌کارگرفته شوند. دو ویژگی تغییر زمان و تغییر مکان رخداد گزارش‌شده در جمله، نسبت به جملهٔ پیشین، امتیازی برای جملهٔ فعلی محسوب می‌شود. همچنین وجود روابط علّی در جمله و نیز شباهت جمله با جستار کاربر امتیاز محسوب می‌شوند که جمله را گزینهٔ مناسبی برای حضور در خلاصه می‌نماید.

۴-۳-۱- ویژگی زمان

در این بخش به تعیین زمان افعال متن، به‌طور خودکار پرداخته و بیش‌تر به این توجه می‌شود که چه جملاتی حاوی تغییرات زمانی نسبت به جملهٔ پیشین هستند. زمان تعریف‌شده برای این پژوهش به سه دستهٔ گذشته، حال و آینده تقسیم شده است. اطلاعات مربوط به زمان از طریق تحلیل صرفی افعال و به‌طور خودکار به‌دست می‌آید؛ برای این منظور از واژگان زایای زبان فارسی که با حمایت شورای عالی اطلاع‌رسانی انجام شده، استفاده گردیده است (اسلامی، ۱۳۸۳). این واژگان حاوی بن ماضی و مضارع تمامی افعال زبان فارسی است که به‌همراه ریشه‌یاب می‌تواند در تعیین زمان فعل و جمله به‌کارگرفته شود. این واژگان همچنین دربرگیرندهٔ افعال کمکی، وجه‌نما و اسنادی به‌ترتیب نظیر "خواهد"، "باید"، "است" می‌باشد که همگی این اطلاعات می‌تواند در تعیین صحیح زمان فعل جمله به‌کار گرفته شوند. به‌عنوان مثال وجود فعل "خواهد" مشخص‌کنندهٔ زمان آینده در ترکیب فعلی است.

۴-۳-۲- ویژگی مکان

در این بخش با استفاده از فهرستی حاوی ۱۴۰۰ اسم مکان شامل اسامی کشورها، شهرهای مهم ایران و دنیا، مکان‌های شاخص ایران و جهان، قاره‌ها و آبهای آزاد و اسامی از این دست و همچنین افعالی که تغییرات مکانی را نشان می‌دهد نظیر "رفتن"، "آوردن"، "بردن" و غیره، این قابلیت را به سامانه دادیم تا بتواند جملات حاوی اطلاعات مکانی را شناسایی کند؛ همچنین جهت بالابردن میزان دقت سامانه در تعیین ویژگی مکان فهرستی از نشان‌گرهای مکانی نظیر "کوچه"، "خیابان"، "کشور" و امثال آن نیز، برای تشخیص خودکار، تهیه شد. درنتیجه سامانه این قابلیت را پیدا می‌کند تا جملات با تغییر مکانی نسبت به جملهٔ پیشین را شناسایی کند.

۴-۳-۳- ویژگی میزان شباهت با جستار[1]

در این سامانه خلاصه‌ساز یک جستار، به‌عنوان ورودی از کاربر دریافت می‌شود. جستار، ترکیبی از کلمات کلیدی مرتبط با مفاهیم موجود در متن است. کاربر با واردکردن جستار از سامانه می‌خواهد که خلاصهٔ تولیدشده را مرتبط با همان جستار تولید کند؛ یعنی خلاصهٔ تولیدشده باید دربرگیرندهٔ رئوسی از مطالب باشد که در جستار مشخص شده است. این ویژگی از طریق محاسبهٔ میزان شباهت (S) کسینوسی هر جمله با کلمات W در عبارت جستار q مشخص می‌شود. هرچه میزان شباهت با جستار برای هر

[1] Query

جمله بیشتر باشد، آن جمله امتیاز بیش‌تری را در این ویژگی کسب می‌کند. فرمول شباهت کسینوسی به‌صورت زیر است، که در آن N تعداد کلمات موجود در جستار می‌باشد:

(۱)

$$S(q_k,d_j)= q_k.d_j = \sum_{i=1}^{N} W_{ik} * W_{ij}$$

مطابق با فرمول (۱)، برای هر جملهٔ موجود در متن اصلی، برداری مبتنی بر بسامد کلمات جستار در جمله ساخته می‌شود. از سوی دیگر نیز برداری مبتنی بر بسامد کلمات جستارداده‌شده توسط کاربر در کل متن ساخته می‌شود. وزن به‌دست آمده برای هر یک از مؤلّفه‌ها بر حسب معیار شناخته شده tf-idf[1] در پیکره محاسبه می‌شود. $d_j=(w_{1j},w_{2j},\ldots,w_{nj})$ و $q_k=(w_{1k},w_{2k},\ldots\ldots w_{nk})$ بردارهای ساخته شده است. در حالت خاصی جستار می‌تواند به‌طور خودکار با توجه به کلمات کلیدی پربسامد استخراج شود. در این صورت کاربر نیازی به وارد کردن جستار ندارد و جستار به‌نوعی شامل هدف اصلی نگارنده از نگارش متن است که در واژگان کلیدی پربسامد منعکس شده است. در بخش ارزیابی این سامانه، عبارت جستار به همین نحو به‌طور خودکار ایجاد شده است.

۴-۳-۴- ویژگی روابط علّی

وجود روابط علّی در جمله بر مبنای نظریهٔ مدل موقعیتی، جمله را از نظر درک متن در جایگاه مهمی قرار می‌دهد. با توجه به اینکه خلاصه‌ساز ایجادشده مبتنی بر چارچوب مدل موقعیتی ایجاد شده است، وجود روابط علّی بین علّت و معلول می‌تواند امتیازی برای به‌کارگیری جمله در خلاصهٔ نهایی باشد. روابط علّی به دو صورت آشکار و تلویحی در متن وجود دارند. در پیاده‌سازی‌های انجام‌شده برای تشخیص روابط علّی، اکثر قریب به اتفاق کارها به تعیین روابط آشکار پرداخته‌اند؛ چون تشخیص روابط تلویحی به‌دلیل پیچیدگی‌های زبانی دشوار است. در زبان فارسی نیز تاکنون کاری در این زمینه انجام نشده است؛ به همین دلیل با مطالعهٔ کارهای انجام‌شده در زبان انگلیسی فهرستی از کلمات که نمایان‌گر وجود یک رابطهٔ علّی در جمله هستند، تهیه و بعد از اطمینان از اینکه همان کارکرد را در زبان فارسی دارند، مورد استفاده قرار گرفتند. از جمله این کلمات می‌توان به کلماتی نظیر "مانع‌شدن"، "باعث‌شدن"، "متعاقباً" و غیره اشاره داشت. به‌عنوان یک کار تکمیلی، مترادف‌های کلمات نشان‌گر روابط علّی، که به‌روش بالا استخراج شدند، از طریق پیمایش رابطه ترادف در فارس‌نت (شمس‌فرد، ۲۰۱۰) به‌دست آمدند و بعد از بررسی در فهرست ایجادشده قرار گرفتند. در نهایت فهرست نود کلمه‌ای از کلمات نشان‌گر روابط علّی، حاصل شد که وقوع آنها در جمله امتیازی از این بابت برای جمله برای به‌کارگیری در خلاصه محسوب می‌شود.

۴-۴- انتخاب جملات

هنگامی که برای هر جمله چهار ویژگی بالا مبتنی بر چارچوب نظری مدل موقعیتی به‌علاوهٔ اطلاعات فاعلی مشخص شد، برای انتخاب جملات، سه روش پیشنهادی معرفی گردید. سه روش پیشنهادی عبارتند از:

۱. انتخاب مهم‌ترین جملات بر حسب امتیازات اخذشده بدون درنظرگرفتن خوشه‌بندی فاعلی
۲. انتخاب جملات با امتیاز بالاتر از مهم‌ترین خوشهٔ فاعلی
۳. انتخاب تعدادی جمله از هر خوشهٔ فاعلی متناسب با حجم هر خوشه و به‌کارگیری آنها در خلاصه

۴-۴-۱- روش نخست: انتخاب مهم‌ترین جملات برحسب امتیاز

در این روش با توجه به ویژگی‌های تغییر زمانی، تغییر مکانی، میزان شباهت با جستار و وجود رابطهٔ علّی امتیازی به هر جمله اختصاص داده می‌شود. بدین ترتیب که اگر تغییر زمانی و تغییر مکانی در جمله دیده شود، به‌ازای هر مورد یک امتیاز و چنانچه رابطهٔ علّی در جمله دیده شود دو امتیاز و میزان شباهت کسینوسی نیز به همان صورت محاسبه‌شده، به‌صورت امتیاز جمله محسوب می‌شود. جملات براساس امتیازات اخذشده مرتب و سپس بدون توجه به خوشه فاعلی، تعدادی جمله با توجه به نرخ مطلوب فشردگی برای ایجاد خلاصه انتخاب می‌شوند.

۴-۴-۲- روش دوم: انتخاب جملات از مهم‌ترین خوشه فاعلی

در این روش به هر جمله‌ای برحسب چهار ویژگی تغییر زمانی، تغییر مکانی، شباهت با جستار و وجود رابطهٔ علّی، امتیازی اختصاص داده می‌شود؛ سپس امتیاز هر خوشه از مجموع امتیازات جملات درون خوشه احتساب می‌شود.

[1] Term frequency-inverse document frequency

بدین ترتیب خوشه‌ای که بیش‌ترین امتیاز را دارد، به‌عنوان مهم‌ترین خوشه شناسایی می‌شود؛ سپس با توجه به نرخ فشردگی، تعدادی جمله، که امتیاز بالاتری در درون مهم‌ترین خوشه داشتند، از این خوشه برای ایجاد خلاصه انتخاب می‌شوند. به‌طور معمول مهم‌ترین خوشه در برگیرندهٔ شخصیت اصلی داستان است که تغییرات مکانی، زمانی و یک تعداد حوادث علت و معلولی، حول آن در مسیری مشخص اتفاق می‌افتند. درحقیقت تعیین مهم‌ترین خوشه می‌تواند تعیین‌کنندهٔ چارچوب تصویر ذهنی ما برای درک متن باشد.

۴-۴-۳- روش سوم: انتخاب تعدادی جمله از هر خوشهٔ فاعلی

روش سوم، بدین صورت است که پس از آنکه جملات براساس فاعل آنها خوشه‌بندی شدند، از هر خوشه متناسب با طول خوشه تعدادی جمله که حائز امتیازات بیشتر هستند، برای به‌کارگیری در ایجاد خلاصه انتخاب می‌شوند. درنهایت جملات انتخاب‌شده از یکی از سه روش ذکرشده در بالا، به‌ترتیبی که در متن اصلی آمده‌اند، به‌صورت خروجی سامانهٔ خلاصه‌ساز در نظر گرفته می‌شوند.

۵- ارزیابی سامانه‌ها

برای ارزیابی سامانه‌ها در ابتدا با استفاده از روش دقت و بازخوانی بهترین سامانه و معماری مناسب آن شناسایی شد. جهت یافتن بهترین روش خلاصه‌سازی از میان روش‌های سه‌گانهٔ ذکرشده در بالا، از روش دقت[1] و بازخوانی[2] و میانگین وزن‌دار استفاده شد. میانگین وزن‌دار ترکیب دو معیار دقت و بازخوانی است که طبق رابطهٔ زیر تعریف می‌شود که در آن P دقت و R بازخوانی هستند.

$$F=2PR/(P+R) \qquad (۲)$$

به‌منظور بررسی سه سامانه، چهل متن انتخاب و خلاصه‌های انسانی با نظارت برای آنها تهیه شد. نتایج این ارزیابی‌ها در جدول (۵-۱) آمده است. بازخوانی برابر است با تعداد جملات مشترک بین خلاصهٔ انسانی و ماشینی، تقسیم بر کل جملات موجود در خلاصهٔ انسانی و دقت برابر است با تعداد جملات مشترک بین خلاصهٔ انسانی و ماشینی، تقسیم بر کل جملات موجود در خلاصهٔ ماشینی (گرنزبکر[3]، ۱۹۹۰). همان‌طور که در جدول (۵-۱) مشخص است، سامانهٔ مبتنی بر دسته‌بندی فاعلی و انتخاب جملات مهم هر خوشهٔ فاعلی دارای عملکرد بهتری نسبت به دو روش دیگر بوده است. به‌نظر می‌رسد این عملکرد بهتر ناشی از حفظ جملات مهم مربوط به کنش‌گرهای مختلف در متن خلاصه باشد. به‌عبارت دیگر در این روش به‌طور هم‌زمان به حفظ تمامی کنش‌گرها و امتیاز جملات توجه شده است.

	دقت	بازخوانی	میانگین همساز
خلاصه‌ساز مبتنی بر روش سوم	۴۸	۴۲	۴۵
خلاصه‌ساز مبتنی بر روش دوم	۴۲	۳۷	۴۰
خلاصه‌ساز مبتنی بر روش اول	۴۱	۳۶	۳۸

(جدول ۵-۱): نتایج ارزیابی دقت و بازخوانی سه روش پیشنهادی

سپس سامانه و معماری برگزیده که همان سامانه و معماری شمارهٔ ۳ بود با سامانهٔ خلاصه‌ساز ایجاز تهیه شده توسط پژوهش‌گران دانشگاه فردوسی مشهد (پورمعصومی، ۱۳۹۳)، مقایسه شد تا عملکرد سامانهٔ پیشنهادی مورد ارزیابی دقیق قرار گیرد. در این آزمایش‌ها نرخ خلاصه‌سازی ۳۰٪ در نظر گرفته شد. نتایج ارزیابی این سامانه به‌همراه نتایج حاصل‌شده از خلاصه‌ساز فارسی ایجاز به چهار داور داده شد تا میزان رضایت کاربران مورد بررسی قرار گیرد. جدول (۵-۲) نتایج بررسی‌های انجام‌شده را به درصد نشان می‌دهد. کیفیت‌های خوب و متوسط و ضعیف که در این جدول اشاره شده، نتایج ارزیابی کیفی خواننده‌ها از خلاصه‌های تولید شده با درنظرگرفتن معیارهای محتوا و انسجام بوده است. خلاصهٔ دارای محتوای مناسب، خلاصه‌ای است که دارای محتوای مهم موجود در متن اصلی باشد و خلاصهٔ منسجم، خلاصه‌ای است که ترتیب آمدن جملات در آن دارای منطق مناسب و مرجع ضمایر نیز در آن مشخص باشد.

	خوب	متوسط	ضعیف
سامانه پیشنهادی	۳۸	۴۸	۱۴
سامانه ایجاز	۴۲	۳۷	۲۱

(جدول ۵- ۲): نتایج ارزیابی دو سامانهٔ پیشنهادی و ایجاز

[1] Precision
[2] Recall
[3] Gernsbacher

۶- بحث و نتیجه‌گیری

میزان دقت و بازخوانی سامانه نشان داد که سامانهٔ سوم که جملات مهم را از تمامی خوشه‌های فاعلی انتخاب می‌کند، عملکرد بهتری نسبت به دو معماری دیگر داشته است؛ بنابراین خوشه‌بندی فاعلی نقش مهمی در گزینش بهینهٔ جملات برای ایجاد یک خلاصهٔ معتبر دارد. بدین ترتیب می‌توان مناسب‌ترین جملات مرتبط با هر یک از شخصیت‌های اصلی را برگزید. یکی از مشکلاتی که برای روش و معماری دوم یعنی انتخاب جملات از مهم‌ترین خوشه می‌تواند وجود داشته باشد، این است که به‌علت استفاده از تنها یک خوشهٔ فاعلی، فقط جملات و کنش‌های مرتبط با یک شخصیت فاعلی در خلاصه وجود خواهد داشت که باعث گسیختگی متن شده و از میزان ادراک متن می‌کاهد. در این روش کنش‌های مربوط به شخصیت‌های تاثیرگذار ولی فرعی برای خلاصه برگزیده نمی‌شوند. در روش نخست انتخاب جملات نامناسب که به‌طور اتفاقی و بیش‌تر به‌دلیل آوردن امتیاز شباهت کسینوسی با جستار حائز نمرهٔ بالایی می‌شدند، از مشکلات سامانه بود. با این وجود، با توجه به عدم نیاز به خوشه‌بندی فاعلی در روش نخست سرعت این روش نسبت به دو روش دیگر بیش‌تر بود. سرعت اجرا در روش‌های دوم و سوم به‌طور تقریبی یکسان بوده‌اند. نتایج به‌دست‌آمده از مقایسهٔ دو سامانه خلاصه‌ساز پیشنهادی و خلاصه‌ساز ایجاز نشان داد که سامانهٔ پیشنهادی عملکرد مناسبی دارد و می‌توان از روش پیشنهادی به‌عنوان روشی مؤثر برای خلاصه‌سازی متون استفاده کرد، البته باید توجه داشت که در معماری سامانهٔ مورد بحث، سعی شد از ویژگی‌های آماری استفاده نشود و امکان بهره‌گیری از نظریه‌های شناختی در طراحی سامانه‌های پردازش زبان طبیعی مورد نظر بود.

اهمیت این پژوهش در این است که برای نخستین‌بار در زبان فارسی تلاش شد تا سامانهٔ خلاصه‌سازی مبتنی بر جنبه‌های شناختی ذهن انسان طراحی شود که از چارچوب نظری شناخته‌شده‌ای در درک متن بهره می‌برد. نتایج نشان داد که می‌توان بدون بهره‌گیری از اطلاعات آماری و تنها با استفاده از یک مدل قطعی مبتنی بر عملکرد ذهن انسان به نتایج خوبی در ایجاد یک سامانهٔ خلاصه‌ساز خودکار دست یافت که با کارهای مشابه در دنیا قابل مقایسه است. هم‌چنین در این ره‌گذر نظریهٔ شناخته‌شدهٔ مدل موقعیتی در درک متن در قالب یک سامانه خلاصه‌ساز در بوتهٔ آزمایش قرار گرفت. نتایج این مطالعه نشان داد این مدل چارچوب مناسبی برای درک عملکرد ذهن ارائه می‌دهد و این مدل می‌تواند دست‌مایهٔ ایجاد سامانه‌های درک متن مبتنی بر عملکرد ذهن انسان باشد.

۷- منابع

اسلامی محرم، شریفی آتشگاه مسعود، علیزاده لمجیری صدیقه، و زندی طاهره، واژگان زایای زبان فارسی. مجموعه مقالات اولین کارگاه پژوهشی زبان فارسی و رایانه. تهران، ۱۳۸۳.

پورمعصومی آصف، کاهانی محسن، طوسی احمد، استیری احمد ، ایجاز: یک سامانهٔ عملیاتی برای خلاصه‌سازی تک‌سندی متون خبری فارسی. دوفصل‌نامه علمی پژوهشی پردازش هوشمند علائم و داده‌ها. ۳۳-۴۸: (۱) ۱۱؛ ۱۳۹۳.۱.

سلطان‌زاده فاطمه، وزیرنژاد بهرام، "یک سامانهٔ خودکار جهت خطایابی متون فارسی"، همایش زبان فارسی و اینترنت، انجمن زبان‌شناسی ایران، آبان ۱۳۹۲.

AleAhmad Abolfazl, Amiri Hadi, Darrudi Ehsan ,Rahgozar Masoud ,Oroumchian Farhad, Hamshahri: A standard Persian text collection, Journal of Knowledge-Based Systems, July 2009, Vol. 22 No.5, p.382-387.

Aries Abdelkerim, Oufaida Houda, and Nouali Omar, Using clustering and a modified classification algorithm for automatic text summarization, in IS&T/-SPIE Electronic Imaging, 2013, pp. 865811–865811.

Bijankhan Mahmood,and etc. Lessons from building a Persian written corpus: Peykare, Language resources and evaluation, 2010, vol. 45 no.2 pp.143–164.

Blanco Eduardo, Castell Nuria, and Moldovan Dan, Causal Relation Extraction., in LREC, 2008.

Carbonell Jaime and Goldstein Jade, The use of MMR, diversity-based reranking for reordering documents and producing summaries, in Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998, pp. 335–336.

Deerwester Scott and etc. , Indexing by latent semantic analysis, JASIS, 1990, vol. 41, no. 6, pp. 391–407 .

Edmundson Harold P., New methods in automatic extracting, J. ACM JACM, 1969, vol. 16, no. 2, pp. 264–285.

Filatova Elena and Hatzivassiloglou Vasileios, A formal model for information selection in multi-sent-

Van Dijk, and etc., Strategies of discourse comprehension. Academic Press New York, 1983.

Wolff Phillip and Song Grace, Models of causation and the semantics of causal verbs, Cognit. Psychol., 2003, vol. 47, no. 3, pp. 276–332.

Zwaan Rolf A., Langston Mark C., and Graesser Arthur, The construction of situation models in narrative comprehension: An event-indexing model, Psychol. Sci., pp. 292–297, 1995.

سیّده ساره صادقی تحصیلات دوره کارشناسی را در دانشگاه شهید چمران اهواز در سال ۱۳۹۰ به اتمام رسانده است. وی هم‌اکنون دانشجوی رشتهٔ کارشناسی ارشد زبان‌شناسی رایانشی دانشگاه شریف است. زمینه‌های پژوهشی مورد علاقهٔ ایشان در حوزه‌های خلاصه‌سازی متن، تهیهٔ پیکره و تحلیل معنایی است.

نشانی رایانامهٔ ایشان عبارت است از:

Sarehsadeghi36@yahoo.com

بهرام وزیرنژاد عضو هیأت علمی دانشگاه صنعتی شریف است. ایشان دکترای خود را در سال ۱۳۸۷ در رشتهٔ مهندسی پزشکی- بیوالکتریک از دانشگاه صنعتی امیرکبیر دریافت کرد. وی طی دو دوره در سال‌های ۲۰۰۸ و ۲۰۱۵ به‌عنوان پژوهش‌گر مهمان در دانشگاه سیدنی استرالیا مشغول به امور پژوهشی و تدریس بود. از او بیش از چهل مقاله در کنفرانس‌ها و نشریات معتبر داخلی و خارجی به چاپ رسیده است. زمینه‌های پژوهشی مورد علاقهٔ ایشان پردازش گفتار، پردازش زبان طبیعی، داده‌کاوی، هوش مصنوعی و زبان‌شناسی رایانشی است.

نشانی رایانامهٔ ایشان عبارت است از:

bahram@sharif.edu

ence text extraction, in Proceedings of the 20th international conference on Computational Linguistics, 2004, p. 397.

Gernsbacher Morton A., Language comprehension as structure building. Routledge, 1990.

Girju Roxana, Automatic detection of causal relations for question answering, in Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering, 2003, Volume 12, pp. 76–83.

Hart Roger A. and Moore Gary T., The development of spatial cognition: A review. AldineTransaction, 1973.

Keefe David, The time course and durability of predictiveinferences, J. Mem. Lang.,1993, vol. 32, no. 4, pp. 446–463.

Lin Hui and Bilmes Jeff, Multi-document summarization via budgeted maximization of submodular functions, in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010, pp. 912–920.

Luhn Hans Peter, The automatic creation of literature abstracts, IBM J. Res. Dev., 1958, vol. 2, no. 2, pp. 159–165.

Mani Inderjeet and Bloedorn Eric, Multi-document summarization by graph search and matching, ArXiv Prepr. Cmp-Lg9712004, 1997.

Miller George, WordNet: a lexical database for English, Commun. ACM, 1995, vol. 38, no. 11, pp. 39–41.

Sauper Christina and Barzilay Regina, Automatically generating wikipedia articles: A structure-aware approach, in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2009, Volume 1-Volume 1, pp. 208–216.

Schiffman Barry, Mani Inderjeet, and Concepcion Kristian, Producing biographical summaries: Combining linguistic knowledge with corpus statistics, in Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, 2001, pp. 458–465.

Shamsfard Mehrnoosh, and etc. Semi Automatic Development of FarsNet; The Persian WordNet, Accepted at Global WordNet Conference (GWA-2010), India, 2010.

Turpin Andrew, Tsegay Yohannes, Hawking David, and Williams Hugh, Fast generation of result snippets in web search, in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 127–134.