سیستم خودکار خلاصه‌ساز متون فارسی

فاطمه شفیعی[۱]، مهرنوش شمس‌فرد[۲]

[۱] آزمایشگاه پردازش زبان طبیعی، دانشکده علوم و مهندسی کامپیوتر، دانشگاه شهید بهشتی
f.shafiee@hotmail.com

[۲] آزمایشگاه پردازش زبان طبیعی، دانشکده علوم و مهندسی کامپیوتر، دانشگاه شهید بهشتی
m-shams@sbu.ac.ir

چکیده

با رشد روزافزون مستندات متنی در وب، انتخاب اطلاعات مطلوب در زمان محدود کار مشکلی است. با استفاده از ابزارهایی نظیر خلاصه‌سازها، می‌توان این حجم انبوه اطلاعات را با تولید خلاصه پیش‌نویس مدیریت نمود. تاکنون رویکردهای متنوعی برای زبان‌های مختلف ارائه شده‌اند که قدمت برخی به شصت سال نیز می‌رسد. در این مقاله، روشی نوین برای خلاصه‌سازی متون خبری فارسی با دقتی بالاتر از سیستم‌های موجود معرفی شده است. این خلاصه‌ساز با بهره‌گیری از دانش موجود در فارس‌نت، جملات را بر اساس میزان شباهت و ارتباط آنها به یکدیگر، خوشه‌بندی می‌نماید. سپس با استفاده از خوشه‌های حاصل، خلاصه نهایی به گونه‌ای تولید می‌گردد که جملات آن دارای کمترین میزان افزونگی و بیشترین ارتباط است. همچنین در صورت نیاز، با بهره‌گیری از خوشه‌های هم‌وقوع، ابهامات موجود در متن خلاصه نیز رفع می‌گردند.

کلمات کلیدی

خلاصه‌سازی استخراجی، خوشه‌بندی، روابط معنایی، شباهت‌سنجی، پردازش زبان طبیعی، فارس‌نت.

۱- مقدمه

یکی از انواع سیستم‌هایی که در تحلیل و پردازش متون وجود دارد، سیستم خلاصه‌ساز اسناد است. خلاصه‌سازی خودکار سند، یعنی تولید یک نسخه مختصرتر از سند اصلی به نحوی که ویژگی‌ها و نکات اصلی سند اولیه حفظ شود [6]. سیستم‌های خلاصه‌ساز در دنیای امروز کاربردهای فراوانی دارند مانند تولید خلاصه‌هایی از اطلاعات بهداشتی درمانی [23] و همچنین همچنین در تولید سیستم‌های تصمیم‌یار [25].

خلاصه سازی ممکن است به صورت استخراجی[۱] یا چکیده سازی[۲] انجام شود. خلاصه‌سازی استخراجی با انتخاب جملاتی از سند ورودی خلاصه را تولید می‌نماید به نحوی که میان جملات خلاصه انسجام و پیوستگی برقرار باشد و همچنین از افزونگی جلوگیری شود. به طور کلی سه مرحله پیش‌پردازش، پردازش و تولید در فرآیند این نوع خلاصه‌سازی وجود دارد. در نخستین مرحله باید متن ورودی به ساختاری قابل پردازش توسط خلاصه‌ساز تبدیل گردد. مشخص نمودن محدوده جملات و کلمات، تشخیص اسامی خاص و مرکب، ریشه‌یابی، حذف ایست‌واژه‌ها[۳]، شناسایی چند-گرام‌ها[۴]، برچسب‌زنی مقوله نحوی[۵] و برچسب‌زنی نقش معنایی[۶] کلمات را می‌توان از جمله وظایف این فاز در نظر گرفت. در مرحله پردازش، جملات مهم شناسایی می‌شوند تا در خلاصه نهایی قرار بگیرند. تاکنون در منابع و مقالات مختلف دسته‌بندی‌های متنوعی برای الگوریتم‌ها و روش‌های شناسایی جملات مهم پیشنهاد شده که در ادامه ترکیبی از این دسته‌بندی‌ها و روش‌های موجود در هر دسته ذکر گردیده است.

در رویکردهای مبتنی بر ویژگی، خلاصه‌ساز با شناسایی ویژگی‌های ظاهری یک جمله، امتیازی را به جمله نسبت می‌دهد. مکان جمله در متن، طول جمله و حضور یا عدم حضور کلمات کلیدی، اسامی خاص، کلمات مندرج در عنوان سند ورودی، ضمایر، کلمات با حروف بزرگ و عبارات کلیدی در متن بخشی از این ویژگی‌های ظاهری هستند که در اغلب خلاصه‌سازهای استخراجی امروزی وجود دارند. از الگوریتم‌های یادگیری ماشین نیز در خلاصه‌سازها استفاده می‌شود، مانند مدل پنهان مارکف [16] و درخت‌های تصمیم [28]. رویکردهای مبتنی بر تحلیل زبان طبیعی بر خلاف الگوریتم‌های

یادگیری ماشین، نیازی به آموزش و یادگیری ندارند. بهره‌گیری از هستان‌شناسی[۷] [24] در خلاصه‌سازی از جمله روش‌های موجود در این رویکرد هستند. دسته‌ای دیگر از خلاصه‌سازها مبتنی بر روش LSA کار می‌کنند [26]. در رویکردهای مبتنی بر گراف سند به گرافی غیرجهت‌دار تبدیل می‌شود که جملات سند گره‌های گراف را تشکیل می‌دهند. جملات مهم، گره‌هایی هستند که بیشترین میزان اتصال را با دیگر گره‌ها دارند [15]. سیستم خلاصه‌ساز Barzilay و Elhadad با استفاده از رویکردی مبتنی بر زنجیره‌های لغوی معرفی شد [27]. در برخی از خلاصه‌سازها از الگوریتم‌های تکاملی [10]، منطق فازی [21] و یا ترکیب این دو رویکرد [7] بهره برده شده است. در رویکردهای مبتنی بر خوشه‌بندی، با استفاده از معیارهای شباهت‌سنجی گوناگون جملات مشابه در یک خوشه قرار می‌گیرند. سپس با انتخاب جملات از این خوشه‌ها، خلاصه نهایی تولید می‌گردد. ایده استفاده از الگوریتم‌های خوشه‌بندی در روش‌های خلاصه‌سازی در حال پیشرفت است و روش‌های متنوعی جهت بهبود این رویکرد معرفی می‌شوند، مانند روش پیشنهادی Xia [9]. با استفاده از ویژگی‌های ظاهری جملات در شباهت‌سنجی و خوشه‌بندی، این رویکردها در شناسایی افزونگی در مجموعه‌ای از اسناد ورودی به خوبی عمل می‌کنند اما با اعمال دانش نیز می‌توان نتایج بهینه‌تری بدست آورد. ما نیز در روش پیشنهادی این مقاله، از اعمال دانش برای خوشه‌بندی جملات استفاده نموده‌ایم.

با وجود گذشت شصت سال از معرفی اولین سیستم خلاصه‌ساز برای زبان انگلیسی هنوز کارهای انجام شده در زمینه خلاصه‌سازی متون فارسی در مقایسه با زبان انگلیسی بسیار اندک می‌باشد. در واقع می‌توان گفت خلاصه‌سازی اسناد فارسی، حیطه‌ای نوین محسوب می‌شود که تعداد کارهای انجام شده در حال افزایش است که در ادامه برخی از این روش‌های پیشنهادی معرفی شده‌اند. شمس‌فرد و کریمی در سال ۱۳۸۵ خلاصه‌ساز تک‌سندی استخراجی را پیشنهاد کرده بودند [۲]. الگوریتم این خلاصه‌ساز مبتنی بر نظریه گراف و زنجیره لغوی می‌باشد. با بهبود خلاصه‌ساز شمس‌فرد و کریمی، خلاصه‌ساز پارسامیست در سال ۱۳۸۷ معرفی شد [۳]. Azom نام سیستم خلاصه‌ساز پیشنهاد شده زمانی‌فر و کاشفی است که بر اساس ویژگی‌های آماری و مفهومی متن و تبدیل ساختار سند ورودی به درخت فراکتال، خلاصه‌ای از آن را تولید می‌نماید [8]. در روش پیشنهادی توفیقی و همکاران نیز ساختار متن به صورت یک درخت فراکتال فرض شده و سپس امتیاز هر جمله بر اساس شش ویژگی ظاهری متن محاسبه می‌شود [19]. روش پیشنهادی شاکری و همکاران نیز مبتنی بر گراف می‌باشد که در آن علاوه بر توجه به میزان وزن هر جمله، به روابط موجود بین هر دو جمله نیز توجه شده است [5]. ایجاز، سیستم خلاصه‌ساز تک‌سندی و چندسندی برای متون فارسی و انگلیسی می‌باشد که متعلق به دانشگاه فردوسی مشهد است [17]. این خلاصه‌سازها روش‌هایی هستند که طی چهار، پنج سال گذشته در زبان فارسی پیشنهاد و معرفی شده‌اند. بعضی از روش‌های خلاصه‌سازی اولیه که نتایج مناسبی را تولید نموده‌اند شامل خلاصه‌ساز فارسی‌سام [14]، روش پیشنهادی شهابی [۱] و همچنین خلاصه‌ساز هنرپیشه [4] هستند.

با وجود نتایج نسبتا مناسب این سیستم‌ها، نیاز به بررسی مجدد روش‌های گوناگون خلاصه‌سازی متون فارسی است تا روشی بهینه‌تر جهت برآورده نمودن انتظارات و توقعات کاربران ارائه گردد. تمرکز این مقاله بر روی خلاصه‌سازی استخراجی متون خبری توسط روشی نوین مبتنی بر خوشه‌بندی جملات به دو نوع خوشه اصلی بر اساس میزان شباهت و ارتباط آنها به یکدیگر با استفاده از دانش موجود در فارس‌نت [11،20] می‌باشد. در بخش دوم این نوشتار، به معرفی روش پیشنهادی پرداخته شده است. بخش سوم به ارزیابی این خلاصه‌ساز و مقایسه آن با خلاصه‌سازهای مشابه اختصاص دارد. در نهایت بخش چهارم دربردارنده نتیجه‌گیری و جمع‌بندی است.

۲- روش پیشنهادی

در روش پیشنهادی ابتدا با استفاده از هشت ویژگی ظاهری از پیش تعیین شده، مقدار امتیاز ویژگی برای تمام جملات سند ورودی محاسبه می شود. سپس جملات در سه دسته خوشه مشابه‌ها، مرتبط‌ها و هم‌وقوع‌ها، خوشه‌بندی می‌شوند. هر یک از خوشه‌های جملات مشابه شامل تعدادی جمله شبیه است. خوشه‌های جملات مرتبط دربردارنده جملاتی هستند که بیشترین ارتباط معنایی را با یکدیگر دارند. خوشه‌های هم وقوع نیز جهت روان‌سازی متن خلاصه در نظر گرفته شده‌اند. متن خلاصه اولیه با استفاده از این خوشه‌ها تولید می‌گردد. با اعمال مقدار ضریب فشردگی بر خلاصه اولیه، متن خلاصه نهایی بدست می‌آید. روش خلاصه‌سازی پیشنهاد شده شامل سه مرحله پیش‌پردازش، پردازش و تولید خلاصه می‌باشد که در ادامه هر مرحله بیان شده است. قابل ذکر می‌باشد که سیستم پیشنهادی به دو صورت تک‌سندی و چندسندی برای متون خبری پیاده‌سازی گردیده به طوری که در حالت چندسندی، با فراهم نمودن چند سند در رابطه با موضوعی مشخص، ابتدا تمام اسناد باید به یکدیگر متصل و سپس به عنوان یک سند ورودی به سیستم داده شود.

۲-۱- مرحله اول - پیش‌پردازش

قطعه‌بندی: اولین گام در مرحله پیش‌پردازش، تشخیص محدوده جملات و کلمات است. در این مرحله از قطعه‌بند STeP1 استفاده شده است [22].

حذف ایست‌واژه‌ها: برای شناسایی و حذف این کلمات، لیست ایست‌واژه‌های موجود در سیستم خلاصه‌ساز پارسامیست مورد استفاده قرار گرفت که تکمیل‌تر شده است. این لیست شامل ضمایر اشاره، حروف ربط، حروف اضافه، افعال ربطی، ضمایر و اکثر افعال ساده و قیود می‌باشد.

شناسایی مقادیر عددی: این کلمات پس از شناسایی، برچسب مقدار عددی دریافت می‌کنند.

شناسایی اسامی خاص: این کلمات نیز با استفاده از یک لیست از پیش فراهم شده قابل شناسایی هستند. این لیست شامل کلماتی مانند نام و فامیل افراد، اسامی کشورها، شهرها، مکان‌ها و رویدادها می‌باشد. تشخیص این کلمات جهت روان‌سازی متن خلاصه، ضروری است زیرا اکثر این اسامی خاص، مراجع ضمایر موجود در جملات بعدی می‌باشند.

ریشه‌یابی: پس از شناسایی ایست‌واژه‌ها، مقادیر عددی و اسامی خاص، کلمات باقی‌مانده وارد فاز ریشه‌یابی می‌شوند. در این مرحله از ریشه‌یاب STeP1 استفاده شده است [22].

آماده‌سازی منابع زبانی: روش پیشنهادی این مقاله از روابط موجود در فارس‌نت بهره برده است. برای هر کلمه موجود در متن اصلی، عنوان و کلمات کلیدی، مجموعه اطلاعات معنایی مورد نیاز از پایگاه داده فارس‌نت استخراج و در سیستم ذخیره می‌شود تا در فاز پردازش از آن‌ها استفاده گردد.

۲-۲- مرحله دوم - پردازش

مرحله پردازش شامل سه عمل اصلی می‌باشد که در ادامه آمده است:

محاسبه امتیاز ویژگی: در این قسمت امتیازدهی جملات متن ورودی با استفاده از هشت ویژگی ظاهری صورت می‌گیرد. در این بخش، هشت ویژگی ظاهری را که در اکثر خلاصه‌سازها وجود دارند انتخاب نمودیم که در ادامه معرفی شده‌اند.

طول جمله: با این فرض که معمولا جملات طولانی‌تر حاوی اطلاعات مهم‌تری هستند، طول جمله به عنوان یک ویژگی در نظر گرفته شده است. برای محاسبه امتیاز طول هر جمله، از رابطه (۱) استفاده می‌شود:

$$F1_{(x)\text{جمله}} = \frac{\text{تعداد کلمات در جمله}(x)}{\text{تعداد کلمات در طولانی‌ترین جمله متن}} \quad (۱)$$

مکان جمله: در اکثر متون، به خصوص متن خبری، جملات اول و انتهایی هر پاراگراف، از جملات میانی آن پاراگراف مهم‌تر هستند. رابطه (۲) مقدار امتیاز این ویژگی را محاسبه می‌نماید:

$$F2_{(x)\text{جمله}} = \text{Max}\left[\frac{1}{(\text{مکان جمله}(x)\text{ در پاراگراف} + 1)}, \frac{1}{\text{تعداد جملات در پاراگراف}}\right] \quad (۲)$$

کلمات اشاره[۸]: کلماتی هستند که در صورت قرار گرفتن آن‌ها در جمله‌ای، جمله مذکور حاوی خبر مهم یا اطلاعات مفیدی است و یا معمولا نشان‌دهنده جمع‌بندی هستند، مانند "بنابراین" و یا "نتیجه" [۳]. رابطه (۳) نحوه محاسبه مقدار امتیاز این ویژگی را نمایش می‌دهد:

$$F3_{(x)\text{جمله}} = \frac{\text{تعداد کلمات اشاره در جمله}(x)}{\text{Max}\left(\text{تعداد کلمات اشاره در جملات موجود در پاراگراف حاوی جمله}(x)\right)} \quad (۳)$$

کلمات موجود در عنوان: جملاتی که حاوی کلمات موجود در عنوان متن می‌باشند، از اهمیت بیشتری برخوردار هستند. رابطه (۴) برای هر جمله مقدار امتیاز این ویژگی را محاسبه می‌نماید:

$$F4_{(x)\text{جمله}} = \frac{\text{تعداد کلمات عنوان موجود در جمله}(x)}{\text{تعداد کلمات در عنوان متن}} \quad (۴)$$

کلمات کلیدی: جملاتی که دربردارنده تعداد کلمات کلیدی بیشتری می‌باشند، از جملات دیگر متن مهم‌تر هستند. منظور از کلمات کلیدی، کلماتی هستند که کاربر می‌تواند به همراه عنوان و متن ورودی، وارد سیستم خلاصه‌ساز نماید. از رابطه (۵) برای محاسبه مقدار امتیاز این ویژگی استفاده شده است:

$$F5_{(x)\text{جمله}} = \frac{\text{تعداد کلمات کلیدی در جمله}(x)}{\text{تعداد کلمات کلیدی}} \quad (۵)$$

وزن کلمه: وزن هر کلمه با استفاده از رابطه (۶)، روش بسامد کلمه - معکوس بسامد سند، قابل محاسبه است. سپس برای هر جمله، با استفاده از رابطه (۷) مقدار امتیاز این ویژگی بدست می‌آید:

$$\text{وزن کلمه (i)} = \text{تعداد تکرار کلمه در جملات متن} \times \log\frac{\text{تعداد جملات متن}}{\text{تعداد جملات حاوی کلمه i}} \quad (۶)$$

$$F6_{(x)\text{جمله}} = \frac{\text{مجموع وزن کلمات جمله}(x)}{\text{Max}\left(\text{مجموع وزن کلمات جملات متن}\right)} \quad (۷)$$

اسامی خاص و مقادیر عددی: اغلب جملاتی که حاوی اسامی خاص و مقادیر عددی هستند، دربردارنده اطلاعات مهمی هستند. برای محاسبه مقدار امتیاز این دو ویژگی از روابط (۸) و (۹) استفاده شده است:

$$F7_{(x)\text{جمله}} = \frac{\text{تعداد اسامی خاص در جمله}(x)}{\text{طول جمله}(x)} \quad (۸)$$

$$F8_{(x)\text{جمله}} = \frac{\text{تعداد مقادیر داده‌ای در جمله}(x)}{\text{طول جمله}(x)} \quad (۹)$$

مقادیر این ویژگی‌ها برای تمام جملات متن ورودی محاسبه می‌شود. سپس امتیاز هر جمله از مجموع وزن‌دار این هشت مقدار بدست می‌آید. وزن مناسب هر ویژگی بدین ترتیب بدست آمد که وزن یک ویژگی را برابر با مقدار یک و وزن هفت ویژگی دیگر را صفر قرار دادیم. با استفاده از بخشی از متون موجود در بخش تک‌سندی و چندسندی پیکره فراهم شده، فرایند خلاصه‌سازی جهت تعیین مقدار بهینه وزن ویژگی مورد نظر (که وزن آن به طور موقت برابر با مقدار یک فرض شده است) انجام گرفت. مقادیر معیار F-measure به ازای هر سند ورودی محاسبه گردید. پس از تکرار این عملیات در هشت مرحله مجزا برای هر هشت ویژگی تعبیه شده در خلاصه‌ساز، وزن‌های نهایی هر ویژگی از میانگین مقادیر معیار F-measure در دو حالت تک‌سندی و چندسندی بدست آمد.

محاسبه امتیاز شباهت و ارتباط برای هر زوج جمله: دو جمله می‌توانند از لحاظ معنا به یکدیگر شبیه و یا مرتبط باشند مانند "او بیمار است" و "آن پسر مریض است" دو جمله شبیه و" او بیمار است" و "او در بیمارستان بستری شد" دو جمله مرتبط هستند. جهت تعیین میزان شباهت یا ارتباط هر زوج جمله، به مجموعه‌ای از اطلاعات معنایی نیاز می‌باشد. از روابط معنایی متنوع موجود بین کلمات می‌توان برای تعیین میزان ارتباط یا شباهت استفاده نمود. به عنوان مثال "اتومبیل" به "قایق" بیشتر شبیه است تا به "درخت" زیرا "اتومبیل" و "قایق" دارای پدر مشترکی به نام "وسایل نقلیه" هستند [29]. همچنین مفاهیم نیز می‌توانند به یکدیگر مرتبط باشند مانند "چرخ" قسمتی از "اتومبیل" است، "برف" از "آب" درست شده است و یا "اتومبیل" نوعی "وسیله نقلیه" است [12]. پس به عنوان نمونه می‌توان از روابط "ابرمفهوم" در محاسبه شباهت و از روابط "مرتبط است با" در محاسبه میزان ارتباط بهره برد. پس از تعیین امتیاز ویژگی هر جمله با استفاده از روابط بخش قبل، میزان شباهت و ارتباط هر دو جمله از متن محاسبه می‌شود. Silveira و Branco از روابط (۱۰)، (۱۱) و (۱۲) جهت تعیین میزان شباهت دو جمله استفاده کرده‌اند [13].

$$\text{subsequences}(s_1, s_2) = \frac{\sum_i\left(\frac{\text{subsequence}_i}{\text{totalWords}_{S1}} + \frac{\text{subsequence}_i}{\text{totalWords}_{s2}}\right)}{\text{totalSubsequences}} \quad (۱۰)$$

در رابطه (۱۰)، subsequences(s1, s2) تعداد کلمات هم‌پوشان[۹] در زیر رشته‌های موجود بین دو جمله، i تعداد زیر رشته‌های موجود بین دو جمله s1 و s2، subsequencei تعداد کلمات در زیر رشته iام، totalWordsSj تعداد کلمات در جمله jام (j=۱و۲) و totalSubsequences تعداد کل زیر رشته‌های موجود بین دو جمله را نشان می‌دهند.

$$overlap(s_1, s_2) = \frac{\sum commonWords(s_1, s_2)}{totalWords_{S1} + totalWords_{S2} - \sum commonWords(s_1, s_2)} \quad (۱۱)$$

در رابطه (۱۱)، overlap(s_1, s_2) تعداد کلمات هم‌پوشان در دو جمله و commonWords(s_1, s_2) تعداد کلمات مشترک بین دو جمله را نشان می‌دهند. رابطه (۱۲)، جهت محاسبه میزان شباهت استفاده شده است:

$$similarity(s_1, s_2) = \frac{subsequences(s_1, s_2) + overlap(s_1, s_2)}{2} \quad (۱۲)$$

در روش پیشنهادی این مقاله، تغییراتی در سه رابطه مذکور اعمال گردیده که در ادامه بیان شده است. در رابطه (۱۰) ابتدا کلمات اصلی متن

ورودی که جزو ایست‌واژه‌ها، مقادیر عددی و اسامی خاص محسوب نمی‌شوند، شناسایی شده و مترادف‌هایشان از فارس‌نت استخراج می‌گردند. سپس جهت یافتن زیر رشته‌ها، هر کلمه به همراه مجموعه کلمات هم‌معنی‌اش به عنوان یک واحد فرض می‌شوند. در رابطه (۱۱) جهت محاسبه تعداد کلمات مشترک بین دو جمله، هر کلمه به همراه مجموعه‌ای از کلمات مشابه و مرتبط گسترش می‌یابد. اگر در رابطه (۱۱)، هر کلمه به همراه مجموعه کلمات مشابه‌اش به عنوان یک واحد فرض شده باشند، آنگاه رابطه (۱۲) نشان‌دهنده مقدار شباهت دو جمله است. در غیر اینصورت اگر در رابطه (۱۱)، هر کلمه به همراه مجموعه کلمات مرتبطش به عنوان یک واحد فرض شده باشند، آنگاه رابطه (۱۲) مقدار ارتباط دو جمله را نشان خواهد داد. بنابراین می‌توان این رابطه را به دو رابطه (۱۳) و (۱۴) تجزیه نمود:

$$similarity\,(s_1,s_2)=\frac{subsequences(s_1,s_2)+\ overlap_{Similar}(s_1,s_2)}{2} \quad (۱۳)$$

$$relatedness\,(s_1,s_2)=\frac{subsequences(s_1,s_2)+\ overlap_{related\ to}(s_1,s_2)}{2} \quad (۱۴)$$

بنابراین با استفاده از این روابط، میزان شباهت و ارتباط هر جمله با جملات دیگر متن محاسبه و ذخیره می‌گردد. سپس از این مقادیر بدست آمده جهت خوشه‌بندی جملات استفاده خواهد شد.

خوشه‌بندی جملات در سه دسته خوشه اصلی حاوی جملات مشابه، جملات مرتبط و جملات هم وقوع: برای خوشه‌بندی جملات مشابه از الگوریتم ساده‌ای استفاده شده است که با استفاده از امتیاز شباهت بدست آمده در مرحله گذشته، جملات را خوشه‌بندی می‌نماید به نحوی که هر خوشه یک جمله یا چند جمله مشابه را در برگیرد. الگوریتم خوشه‌بندی جملات مرتبط نیز با بهره‌گیری از مقادیر میزان ارتباط، جملات را خوشه‌بندی می‌کند. بنابراین هر جمله متن اصلی، عضو یا مرکز خوشه‌ای است که آن خوشه دربردارنده مجموعه‌ای از مرتبط‌ترین و یا مشابه‌ترین جملات به یکدیگر است. گاه جملاتی در متون مشاهده می‌شوند که حضور آن‌ها، باعث ایجاد ابهام در متن می‌گردد مانند جمله حاوی ضمیر در حالی که مرجع ضمیر مبهم هست. برای رفع این ابهام نیاز است، جمله حاوی مرجع ضمیر به متن خلاصه نیز افزوده گردد. الگوریتم تولید خوشه‌های هم‌وقوع، هر جمله متن ورودی را برای حضور یا عدم حضور این گونه کلمات بررسی می‌نماید. بنابراین جملات موجود در هر خوشه این دسته، باید با هم در متن خلاصه نهایی ظاهر شوند.

۳-۲- مرحله سوم – تولید خلاصه

برای تولید خلاصه اولیه، گزینش جملات از مرکز اولین خوشه حاوی جملات مشابه آغاز می‌گردد. سپس در میان خوشه‌های حاوی جملات مرتبط جستجویی انجام می‌گیرد تا خوشه شامل این جمله و جملات مرتبط با آن شناسایی گردد. اگر مرکز خوشه یافته شده همان جمله کاندیدا باشد، آن‌گاه جمله‌ای از این خوشه گزینش می‌شود که تعداد جملات کمتری از خلاصه در خوشه جملات مشابه‌اش قرار گرفته باشند. اگر چند جمله از خوشه یافته شده با این شرایط کاندیدا گزینش شوند آنگاه جمله‌ای که دارای بیشترین مقدار امتیاز ویژگی است، انتخاب می‌گردد. اما اگر جمله کاندیدا در مرکز خوشه یافته شده قرار نداشته باشد، مرکز خوشه در صورتی در خلاصه درج می‌گردد که دو سوم جملات موجود در متن خلاصه در خوشه جملات مشابه‌اش قرار نداشته باشند. اگر این شرط برای مرکز خوشه برقرار نشود، همانند آنچه بیان شد جمله‌ای از خوشه یافته شده گزینش می‌شود که تعداد جملات کمتری از خلاصه در خوشه جملات مشابه‌اش قرار گرفته باشند. در ادامه الگوریتم، از میان خوشه‌های حاوی جملات مشابه، اگر جمله واقع در مرکز خوشه دوم در خلاصه اولیه وجود نداشته باشد (ممکن است در مرحله قبل، به عنوان جمله مرتبط با جمله اول در خلاصه قرار گرفته باشد)، به همراه مرتبط‌ترین جمله‌اش، در خلاصه اولیه درج می‌شوند. این الگوریتم تا زمانی ادامه می‌یابد که مراکز تمام خوشه‌های حاوی جملات مشابه مورد بررسی قرار گیرند. در انتها، خلاصه اولیه‌ای تولید می‌گردد که کمترین میزان افزونگی و بیشترین ارتباط و پیوستگی را دارا است. سپس جملاتی از متن خلاصه اولیه که در خوشه‌های هم‌وقوع واقع هستند، شناسایی شده و جملات ضروری جهت رفع ابهام از این خوشه‌ها استخراج و به متن خلاصه اولیه نیز افزوده می‌شوند.

در گام بعد، کاربر با وارد نمودن مقدار ضریب فشردگی دلخواه، سیستم را در جهت تولید خلاصه نهایی با طول معین هدایت می‌نماید. با اعمال مقدار ضریب فشردگی بر متن ورودی، طول خلاصه نهایی بدست می‌آید. اگر طول خلاصه نهایی برابر با طول خلاصه اولیه باشد، همان خلاصه اولیه به عنوان خروجی نهایی سیستم به کاربر تحویل داده می‌شود. اگر طول خلاصه اولیه از طول خلاصه نهایی کمتر باشد، از جملات متن ورودی، جملاتی جهت درج در خلاصه نهایی انتخاب می‌شوند که کمترین تعداد جملات مشابه و سپس بیشترین تعداد جملات مرتبط را با جملات خلاصه داشته باشد. اگر طول خلاصه اولیه بیشتر از طول خلاصه نهایی باشد، از میان جملاتی که در خلاصه اولیه درج شده‌اند، جملاتی حذف می‌شوند که بیشترین تعداد جملات مشابه و سپس کمترین تعداد جملات مرتبط را با جملات خلاصه دارد. روند افزایش یا حذف جملات تا زمانی که طول خلاصه اولیه به اندازه طول خلاصه نهایی برسد، ادامه می‌یابد. ممکن است سیستم در حین عملیات حذف یا افزودن، به جای یک جمله، با چند جمله که تعداد جملات مشابه و مرتبط یکسانی دارند برخورد نماید. در این شرایط برای اجرای عمل حذف، جمله‌ای که مقدار امتیاز ویژگی کمتری داشته باشد از متن خلاصه اولیه حذف می‌گردد و برای اجرای عمل اضافه نمودن، جمله‌ای که مقدار امتیاز ویژگی بیشتری داشته باشد به متن خلاصه اولیه افزوده می‌شود. منظور از تعداد جملات مشابه یا مرتبط یک جمله، جملاتی هستند که با آن جمله در یک خوشه حاوی جملات مشابه و یا مرتبط قرار دارند. بنابراین سیستم خلاصه‌ساز با استفاده از این روش قادر به تولید خلاصه نهایی خواهد بود.

۳- ارزیابی روش پیشنهادی

برای ارزیابی خلاصه‌ساز پیشنهادی پنج معیار دقت[۱۰]، فراخوانی[۱۱]، measure-F، ROUGE-L و ROUGE-N انتخاب شده‌اند. این روش خلاصه‌سازی بر روی دو پیکره شامل متون خبری تست گردید که پیکره اول پاسخ نام دارد [18] و پیکره دوم به صورت دستی تهیه شده است.

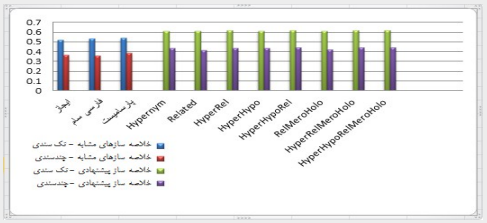
جدول (۱) : مشخصات پیکره‌ها

پیکره	تک‌سندی	چندسندی
اول	۱۰۰ سند با موضوعات مختلف	۵۰ موضوع – هر موضوع حاوی ۲۰ سند
	تعداد خلاصه‌های انسانی هر سند:۵	تعداد خلاصه‌های انسانی هر موضوع:۵
دوم	۴۵ سند با موضوعات مختلف	۳۰ موضوع – هر موضوع حاوی ۶ سند
	تعداد خلاصه‌های انسانی هر سند:۵	تعداد خلاصه‌های انسانی هر موضوع:۵

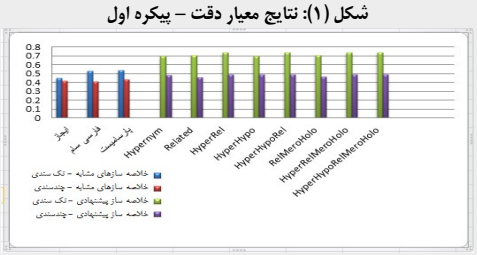
جهت مقایسه و ارزیابی نتایج حاصل از روش پیشنهادی، سه خلاصه‌ساز پارسامیست، فارسی‌سام و ایجاز انتخاب شده‌اند.

۳-۱- آزمایش اول

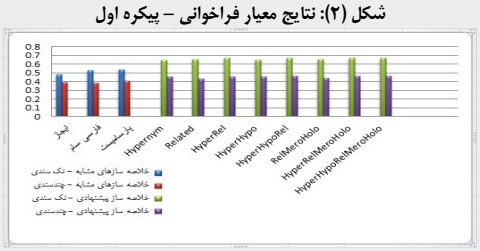
هدف از این آزمایش یافتن بهترین ترکیب روابط معنایی موجود در فارس‌نت در خوشه‌بندی جملات مشابه و مرتبط می‌باشد. دو رابطه "ابرمفهوم" و "مرتبط است با" به عنوان روابط پایه به ترتیب برای تعیین میزان شباهت و ارتباط جملات تعیین شده‌اند. با استفاده از پیکره اول، تاثیر ترکیبات متفاوت این دو رابطه به همراه روابط معنایی دیگر تست گردید تا بهترین روابط معنایی برای شباهت‌سنجی و ارتباط‌سنجی شناسایی شوند. نتایج پنج معیار در ادامه قابل مشاهده هستند.



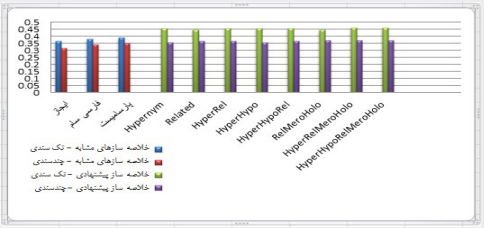
شکل (۱): نتایج معیار دقت – پیکره اول



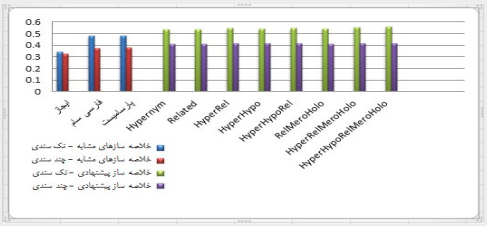
شکل (۲): نتایج معیار فراخوانی – پیکره اول



شکل (۳): نتایج امتیاز F – پیکره اول



شکل (۴): نتایج معیار ROUGE-L – پیکره اول



شکل (۵): نتایج معیار ROUGE-1 – پیکره اول

با توجه به نمودارها، بهترین ترکیب روابط معنایی در خوشه‌بندی بدین صورت است که خلاصه‌ساز از روابط ابرمفهوم[۱۲] و زیرمفهوم[۱۳] برای شباهت‌سنجی و از روابط مرتبط است با[۱۴]، جزءواژگی[۱۵] و کل‌واژگی[۱۶] برای ارتباط‌سنجی استفاده نماید.

۳-۲- آزمایش دوم

هدف از این آزمایش بررسی تاثیر اعمال خوشه‌بندی جملات مشابه با استفاده از پنج رابطه معنایی (خوشه‌بندی نوع اول) بر فرایند خلاصه‌سازی و سپس اضافه نمودن خوشه‌بندی جملات مرتبط (خوشه‌بندی نوع دوم) همانند روال عادی خلاصه‌ساز پیشنهادی است. در خوشه‌بندی‌های این آزمایش از بهترین ترکیب بدست آمده در آزمایش اول استفاده شده است. نتایج حاصل در جدول (۲) و (۳) مشاهده می‌شوند. نتایج حاصل، از تاثیر مثبت خوشه‌بندی جملات مشابه و مرتبط در فرایند خلاصه‌سازی حکایت دارد.

جدول (۲) : مقادیر معیارهای دقت - فراخوانی - F-measure

خوشه‌بندی	دقت		فراخوانی		F-measure	
	تک‌سندی	چندسندی	تک‌سندی	چندسندی	تک‌سندی	چندسندی
نوع اول	۰.۵۷۶۴	۰.۴۲۶۷	۰.۵۸۶۳	۰.۴۵۴۵	۰.۵۸۱۳	۰.۴۴۰۲
نوع دوم	۰.۶۹۵۲	۰.۵۰۳۱	۰.۷۶۴۲	۰.۵۳۱۶	۰.۷۲۸۱	۰.۵۱۷

جدول (۳) : مقادیر معیار ROUGE-L - ROUGE-1

خوشه‌بندی	ROUGE-L		ROUGE-1	
	تک‌سندی	چندسندی	تک‌سندی	چندسندی
نوع اول	۰.۵۰۶۳	۰.۳۷۲۵	۰.۵۱۰۷	۰.۳۹۲۵
نوع دوم	۰.۶۱۸۲	۰.۴۵۳۳	۰.۶۲۲۳	۰.۴۶۸۱

با توجه به جداول، می‌توان این نتیجه را گرفت که وجود خوشه حاوی جملات مرتبط در کنار خوشه حاوی جملات مشابه، برخلاف بسیاری از خلاصه‌سازهای دیگر که فقط بر مبنای شباهت‌سنجی کار می‌کنند، تاثیر مثبتی بر عملکرد خلاصه‌ساز خواهد داشت.

۴- نتیجه‌گیری

در این مقاله خلاصه‌سازی معرفی گردید که با استفاده از از فارس‌نت و روابط معنایی آن، خلاصه‌ای روان، منجسم و با کمترین مقدار افزونگی را تولید می‌نماید. برای تکمیل نمودن این روش پیشنهادی، می‌توان به جای استفاده از کلمات، جملات موجود در مثال‌ها[۱۷] و توضیحات[۱۸] را در روابط شباهت‌سنجی و ارتباط‌سنجی جهت تعیین میزان شباهت و ارتباط اعمال نمود. همچنین برای بهبود عملکرد خلاصه‌ساز می‌توان آن را به نحوی تغییر داد که به جای انتخاب کل جمله، پس از شناسایی عبارات کلیدی و حذف بخش‌های غیرضروری، آن را در خلاصه درج نماید. در روشی دیگر می‌توان از تمام پنج رابطه انتخاب شده از فارس‌نت در خوشه‌بندی جملات مشابه و مرتبط استفاده نمود اما به هر رابطه معنایی در خوشه‌بندی جملات وزنی را نسبت داد.

مراجع

[۱] شهابی امیرشهاب، **"چکیده‌سازی متون زبان فارسی"**، دومین کنفرانس بین‌المللی علوم شناختی، صفحه ۵۶، تهران، ایران، ۱۳۸۱.

[۲] کریمی زهره، شمس‌فرد مهرنوش، **"سیستم خلاصه‌سازی خودکار متون فارسی"**، دوازدهمین کنفرانس بین المللی انجمن کامپیوتر، تهران، ایران، ۱۳۸۵.

[۳] شمس‌فرد مهرنوش، اخوان تارا، عرفانی جورابچی مونا، "**PARSUMIST: خلاصه‌ساز تک‌سندی و چندسندی متون فارسی**"، چهاردهمین کنفرانس بین المللی انجمن کامپیوتر، تهران، ایران، ۱۳۸۷.

[4] Honarpisheh, M. A., Ghassem-Sani, G., Mirroshandel, S. A., *"A Multi-Document Multi-Lingual Automatic Summarization System"*, Proceedings of the 3rd International Joint Conference on natural language processing(IJCNLP), pp.733-738, 2008.

[5] Shakeri, H., Gholamrezazadeh, S., Salehi, M. A., Ghadamyari, F., *"A New Graph-Based Algorithm for Persian Text Summarization"*, Computer Science and Convergence, pp. 21-30, Springer Netherlands, 2012.

[6] Nenkova, A., McKeown, K., *"A survey of text summarization techniques"*, Mining Text Data, pp. 43-76, Springer US, 2012.

[7] Ladekar, A., Mujumdar, A., Nipane, P., Titar, S., Kavitha, M., *"Automatic Text Summarization Using: Fuzzy GA-GP"*, International Journal of Engineering Research and Applications (IJERA), Vol. 2, pp. 1551-1555, 2012.

[8] Zamanifar, A., Kashefi, O., *"AZOM: a Persian structured text summarizer"*, Natural Language Processing and Information Systems, pp. 234-237. Springer Berlin Heidelberg, 2011.

[9] Xia, Y., Zhang, Y., Yao, J., *"Co-clustering sentences and terms for multi-document summarization"*, Computational Linguistics and Intelligent Text Processing, pp. 339-352, Springer Berlin Heidelberg, 2011.

[10] Bossard, A., Rodrigues, C., *"Combining a Multi-Document Update Summarization System –CBSEAS– with a Genetic Algorithm"*, CIMA 2010 - International Workshop on Combinations of Intelligend Methods and Applications, pp. 71-87, France, 2010.

[11] Shamsfard, M., *"Developing FarsNet: A Lexical Ontology for Persian"*, proceedings of the 4th global WordNet conference, 2008.

[12] Banerjee, S., Pedersen, T., *"Extended gloss overlaps as a measure of semantic relatedness"*, IJCAI, Vol. 3, pp. 805-810. 2003.

[13] Silveira, S. B., Branco, A., *"Extracting multi-document summaries with a double clustering approach"*, Natural Language Processing and Information Systems, pp. 70-81,Springer Berlin Heidelberg, 2012.

[14] Hassel, M., Mazdak, N., *"FarsiSum-a persian text summarizer"*, Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, pp. 82-84, 2004.

[15] Erkan, G., Radev, D. R., *"LexRank: Graphbased lexical centrality as salience in text summarization"*, JAIR, Vol. 22, No.1, pp. 457-479, 2004.

[16] Fung, P., Ngai, G., *"One story, one flow: Hidden markov story models for multilingual multidocument summarization"*, ACM Transactions on Speech and Language Processing, Vol. 3, No.2, pp. 1-16, 2006.

[17] Pourmasoumi, A., Kahani, Toosi, S. A., Estiri, A., Qaeim, H., *"Paper: IJAZ: An Operational System for Single-document Summarization of Persian News Texts"*.

[18] Behmadi Moghaddas, B., Kahani, M., Toosi, S. A., Pourmasoumi, A., Estiri, A., *"Pasokh: A standard corpus for the evaluation of Persian text summarizers"*, Computer and Knowledge Engineering (ICCKE), 2013 3th International eConference on, pp. 471-475, 2013.

[19] Tofighy, M., Kashefi, O., Zamanifar, A., Javadi, H. H. S., *"Persian Text Summarization Using Fractal Theory"*, Informatics Engineering and Information Science, pp. 651-662. Springer Berlin Heidelberg, 2011.

[20] Shamsfard, M., Hesabi, A., Fadaei, H., Mansoory, N., Famian, A., Bagherbeigi, S., Assi, S. M., *"Semiautomatic development of farsnet; the persian wordnet"*, Proceedings of 5th Global WordNet Conference, India, 2010.

[21] Suanmali, L., Binwahlan, M. S., Salim, N., *"Sentence Features Fusion for Text summarization using Fuzzy Logic"*, IEEE International Conference on Web Intelligence and Intelligent Agent Technology, Vol. 1, pp. 142-145, 2009.

[22] Shamsfard, M., Jafari, H. S., Ilbeygi, M., *"STeP-1: A Set of Fundamental Tools for Persian Text Processing"*, LREC, 2010.

[23] Feblowitz, J. C., Wright, A., Singh, H., Samal, L., Sittig, D. F., *"Summarization of clinical information: A conceptual model"*, Journal of biomedical informatics 44, No. 4, pp. 688-699, 2011.

[24] Kallimani, J. S., Srinivasa, K. G., Eswara Reddy, B., *"Summarizing News Paper Articles: Experiments with Ontology-Based, Customized, Extractive Text Summary and Word Scoring"*, Cybernetics and Information Technologies 12, No.2, pp. 34-50, 2012.

[25] Workman, T. E., Fiszman, M., Hurdle, J. F., *"Text summarization as a decision support aid"*, BMC Medical Informatics and Decision Making 12, No. 1, 2012.

[26] Ozsoy, M. G., Cicekli, I., Alpaslan, F. N., *"Text summarization of Turkish texts using latent semantic analysis"*, Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), pp. 869-876, 2010.

[27] Barzilay, R., Elhadad, M., *"Text summarizations with lexical Chains"*, Advances in Automatic Text Summarization, pp. 111-121. MIT Press, 1999.

[28] Chen, Y. L., Hung, L. T. H., *"Using decision trees to summarize associative classification rules"*, Expert Systems with Applications: An International Journal, Vol. 36, No.2, pp. 2338-2351, 2009.

[29] Pedersen, T., Patwardhan, S., Michelizzi, J., *"WordNet:: Similarity: measuring the relatedness of concepts"*, Demonstration Papers at HLT-NAACL 2004, pp. 38-41, 2004.

زیرنویس‌ها

1 Extractive
2 Abstractive
3 Stop Words
4 N-grams
5 Part Of Speech Tagging(POS)
6 Semantic Role Labeling(SRL)
7 Ontology
8 Cue Words
9 Overlapping Words
10 Precision Measure
11 Recall Measure
12 Hypernymy
13 Hyponymy(Is-A)
14 Related to
15 Meronymy(Part of)
16 Holonymy(Has-A)
17 Examples
18 Glosses