کارگاه‌های آموزشی مرکز اطلاعات علمی

آموزش مهارت های کاربردی در تدوین و چاپ مقاله

اصول تنظیم قراردادها

مقاله نویسی علوم انسانی

خلاصه‌سازی چندسندی متون فارسی با استفاده از یک روش مبتنی بر خوشه‌بندی

محسن مشکی، مرتضی آنالویی

دانشکده کامپیوتر دانشگاه علم و صنعت ایران

E-mail: mo_meshki@comp.iust.ac.ir, analoui@iust.ac.ir

چکیده - در این مقاله، یک روش جدید مبتنی بر خوشه‌بندی برای خلاصه‌سازی چندسندی متون فارسی پیشنهاد شد. در این روش، پس از پیش پردازش متن شامل تعیین مرز واژه‌ها و جمله‌ها، یکسان‌سازی متن، حذف واژه‌های عمومی و شناسایی عناصر متنی چندتایی، فرآیند اصلی خلاصه‌سازی آغاز می‌شود. در مرحله‌ی خلاصه‌سازی، ابتدا جمله‌ها خوشه‌بندی می‌شود و سپس به ازای هر خوشه جمله‌ای که بیشترین ارتباط با سایر جمله‌ها را دارد، گزینش می‌شود. در آخرین مرحله‌ی خلاصه‌سازی، جمله‌ها با توجه به ترتیب زمانی متن‌ها (خبری) در خلاصه‌ی نهایی درج می‌شوند. نتایج پیاده‌سازی نشان می‌دهند که در بیشتر موارد خروجی سامانه‌ی خلاصه‌سازی پیشنهادی خلاصه‌ی قابل قبولی را تولید می‌کند (بیش از ۸۰ درصد).

کلید واژه- خلاصه‌سازی چندسندی، پیش‌پردازش، خوشه‌بندی، عنصر متنی چندتایی.

۱- مقدمه

با افزایش روز افزون منابع متنی در شبکه جهانی وب، هر روز بر گستره‌ی اطلاعات قابل دسترس برای کاربران افزوده می‌شود. اگر چه این رشد سریع مزایای زیادی دارد، اما مشکلاتی را نیز به همراه دارد. یکی از این مشکلات، سردرگمی کاربران در گزینش مطالب است. هنگامی که یک کاربر پرس‌وجویی را مطرح می‌کند، با سیاهه‌ای از عنوان‌ها مواجه می‌شود و باید زمان زیادی را صرف خواندن آنها برای یافتن اطلاعات مورد نیاز خود کند. سامانه خلاصه‌سازی خودکار متن، یک راه حل برای این مشکل است. در [۱]، سه مزیت عمده برای تولید خودکار خلاصه به وسیله ماشین برشمرده شده است، که عبارتند از:

- اندازه‌ی خلاصه قابل کنترل است، یعنی ماشین می‌تواند خلاصه را با توجه به میزان فشردگی مورد نظر کاربر تهیه کند.
- محتوای آن قابل پیش‌بینی است.
- می‌توان مشخص کرد که هر بخش از خلاصه مربوط به کدام بخش یا بخش‌ها از متن اصلی است.

خلاصه‌سازی یکی از کاربردهای پردازش متن است. پردازش متن شامل چهار سطح است [۲]: پردازش لغوی، پردازش ساختواژی، پردازش نحوی و پردازش معنایی. هر یک از کاربردهای فراوان پردازش متن، از جمله بازیابی اطلاعات، خلاصه‌سازی، درک، تولید، ترجمه، پرسش و پاسخ، استخراج دانش از متون و موارد دیگر با توجه به گستردگی و پیچیدگی، در یک یا چند سطح فوق به انجام می‌رسد.

خلاصه‌سازی به خصوص از نوع چکیده‌سازی، یکی از پیچیده‌ترین کاربردهای پردازش متن است و معمولا با چند سطح از پردازش متن درگیر است. پردازش لغوی یکی از مولفه‌های جدانشدنی خلاصه‌سازی است و در عملیات پیش‌پردازشی به طور گسترده مورد استفاده قرار می‌گیرد که از آن جمله، می‌توان به تعیین مرز

لغات و واژه‌ها اشاره نمود. بکارگیری سایر انواع پردازش، بستگی مستقیم به نوع روش خلاصه‌سازی دارد؛ اما به طور کلی، در روش‌های خلاصه‌سازی گزینشی کمتر و در روش‌های چکیده‌سازی بیشتر مورد استفاده قرار می‌گیرند، به طوریکه در بعضی روش‌های خلاصه‌سازی گزینشی فقط از چند سطح پردازشی محدود و در بعضی دیگر از هر چهار نوع پردازش استفاده می‌شود.

تجربه نشان می‌دهد که اگر چه استفاده از سطوح پردازشی گوناگون تضمینی برای کیفیت خلاصه‌سازی نیست، اما بدون استفاده از سطوح بالای پردازش متن (پردازش‌های نحوی و معنایی) کارایی روش محدود می‌شود. به عنوان مثال: واژه‌ی «کرم» را می‌توان به صورت‌های «کِرم»، «کَرَم»، «کِرِم» و «کُرُم» خواند و یک فارسی زبان با توجه به محتوی جمله و متن قادر به تشخیص آن است.

در سال‌های اخیر فعالیت گسترده‌ای روی ساخت و توسعه‌ی سامانه‌های خودکار خلاصه‌سازی متن برای زبان‌های مختلف انجام شده است. فلسفه‌ی خلاصه‌سازی چند سندی آن است که اطلاعات متنی خاصیت توزیع‌شده و پراکنده دارند. برای گردآوری آنها نیاز به سامانه‌هایی جهت یکپارچه‌سازی آنها با گزینش اطلاعات مفید، حذف بخش‌های افزونه و مرتب نمودن آنها با یک ترتیب منطقی است.

یکی از مهم‌ترین تقسیم‌بندی‌های موجود در مورد خلاصه‌ها، تقسیم آنها به دو نوع گزیده و چکیده است. در این دسته‌بندی، خلاصه‌ها با توجه به منشا متن خروجی به دو نوع گزیده (Extract) و چکیده (Abstract) تقسیم می‌شود. در گزیده، هر جمله از متن به طور مستقیم از متن یا متون اولیه رونوشت می‌شود. از آنجا که در تولید این دسته از خلاصه‌ها، جملات متن تغییرات نحوی و معنایی ندارند، می‌توان آن را نوعی گزینش جملات قلمداد کرد.

از سوی دیگر، چکیده خلاصه‌ای است که دست کم بخش‌هایی از آن در متن یا متون اولیه وجود ندارد. در تولید چکیده، پاره‌ای از جمله‌ها یا همه‌ی آنها بازنویسی می‌شوند. به عنوان مثال، اگر در یک خلاصه جمله «او سیب، انگور و گیلاس‌ها را خورد» به صورت «او میوه‌ها را خورد» نوشته شود، آن خلاصه یک چکیده است.

اگر چه شباهت‌های زیادی بین روش‌های خلاصه‌سازی تک‌سندی و چندسندی وجود دارد، اما تفاوت قابل توجهی بین آنها وجود دارد. چند تفاوت عمده بین خلاصه‌سازی تک‌سندی و چندسندی عبارتند از:

۱. میزان افزونگی اطلاعات در یک دسته از متن‌هایی که به لحاظ موضوعی به هم مربوطند، بیشتر از میزان افزونگی موجود در یک متن است. در خلاصه‌سازی چندسندی، متن‌های ورودی دارای زمینه‌ی مشترکی هستند و گاهی با هدف روشن ساختن موضوع مشابهی نگارش شده‌اند. این تفاوت بین خلاصه‌سازی تک‌سندی و چندسندی، ایجاب می‌کند که برای خلاصه‌سازی چندسندی از روش‌هایی استفاده شود که ضدافزونگی (Anti-Redundancy) هستند.

۲. گاهی در خلاصه‌سازی چندسندی با متن‌هایی مواجه هستیم که دارای یک ترتیب زمانی هستند. به عنوان مثال در خلاصه‌سازی اخبار، هر خبر دارای یک زمان مشخص است و نادیده گرفتن آن در خلاصه‌سازی می‌تواند کیفیت خروجی را پایین بیاورد. در مواردی که متن‌ها دارای یک ترتیب زمانی هستند باید از روش‌هایی سود جست که وزن بیشتری به خبرهای جدید بدهند.

۳. در خلاصه‌سازی چندسندی، نرخ خلاصه‌سازی به طور چشمگیری کمتر از خلاصه‌سازی تک‌سندی است. این تفاوت از آنجا ناشی می‌شود که در خلاصه‌سازی چندسندی با حجم بیشتری از متن‌ها مواجه هستیم و همچنان انتظار می‌رود که خلاصه‌ساز یک خلاصه‌ی کوچک و مفید تولید کند. هر چه تقاضا برای فشردگی بیشتر شود، فرآیند

خلاصه‌سازی سخت‌تر می‌شود.

۴. از آنجا که در خلاصه‌سازی چندسندی، جمله‌های گزینش شده از یک سند نیستند، استفاده از ترتیب اولیه‌ی جمله‌ها در سند اصلی مقدور نیست. در مجموع، مرتب‌سازی جمله‌های گزینش شده برای تولید خلاصه‌ی نهایی در نوع چندسندی بسیار مشکل‌تر از نوع تک‌سندی است.

در این مقاله، یک روش مبتنی بر خوشه‌بندی برای خلاصه‌سازی چندسندی متون فارسی ارایه شده است که خلاصه‌های تولید شده توسط آن از نوع گزیده هستند. در این روش، پس از پیش پردازش متن شامل تعیین مرز واژه‌ها و جمله‌ها، یکسان‌سازی متن، حذف واژه‌های عمومی و شناسایی عناصر متنی چندتایی، فرآیند اصلی خلاصه‌سازی آغاز می‌شود. در مرحله‌ی خلاصه‌سازی، ابتدا جمله‌ها خوشه‌بندی می‌شود و سپس به ازای هر خوشه جمله‌ای که بیشترین ارتباط با سایر جمله‌ها را دارد، گزینش می‌شود. در آخرین مرحله‌ی خلاصه‌سازی، جمله‌ها با توجه به ترتیب زمانی متن‌ها (خبری) در خلاصه‌ی نهایی درج می‌شوند. معیاری شباهتی که در خوشه‌بندی استفاده شده، یک معیار جدید است که از بردارهای واژه-محتوی بهره می‌گیرد.

۲- مروری بر کارهای گذشته

کارهای انجام شده در حوزه‌ی خلاصه‌سازی متون فارسی بسیار اندک و انگشت شمار هستند که بیشتر آنها به خلاصه‌سازی تک‌سندی پرداخته‌اند. در [۳]، یک سامانه‌ی خلاصه‌سازی به نام FarsiSum معرفی شده است. این سامانه، نسخه‌ی تغییر یافته‌ی یک سامانه خلاصه‌سازی متون سوئدی به نام SweSum [۴] برای پوشش زبان فارسی است. خلاصه‌ی خروجی SweSum از نوع گزینشی است و برای زبان‌های سوئدی، نروژی، دانمارکی، اسپانیایی، انگلیسی، فرانسوی و آلمانی پیاده‌سازی شده است. این سامانه خلاصه‌ساز، متن را در قالب متنی یا HTML دریافت می‌کند. متن ورودی می‌تواند از روزنامه انتخاب شود یا به صورت یک گزارش باشد.

در [۵] نیز یک روش خلاصه‌سازی تک‌سندی دیگر پیشنهاد شده است. این سامانه، مانند FarsiSum بر مبنای گزینش جمله‌ها کار می‌کند. همچنین، محتوی خلاصه می‌تواند کلی یا بر اساس پرس‌وجوی کاربر باشد. ایده‌ی بکار رفته در گزینش جمله‌ها در این خلاصه‌ساز، ترکیبی از دو روش زنجیره‌ی لغوی و نظریه‌ی گراف است.

در [۶]، یک روش خلاصه‌سازی چندسندی معرفی شده است که نوع خروجی تولید شده توسط آن از نوع چکیده است. در [۷]، علاوه بر مروری بر روش‌های خلاصه‌سازی موجود، گزارشی از پیاده‌سازی یک نمونه‌ی عملی برای خلاصه‌سازی فارسی ارایه شده است. در [۸]، پس از بررسی موضوعات و چالش‌های مربوط به پردازش متن فارسی، مروری بر روش‌های خلاصه‌سازی موجود صورت گرفته است.

۳- پیش‌پردازش

قبل شروع مراحل اصلی خلاصه‌سازی، نیاز است که متن مورد پیش پردازش قرار گیرد. اعمالی پیش‌پردازشی شامل تعیین مرز واژه‌ها و جمله‌ها، یکسان‌سازی متن، حذف واژه‌های عمومی و شناسایی عناصر متنی چندتایی است. در ادامه، هر یک از این مراحل مورد بررسی قرار خواهند گرفت.

۳-۱- تعیین مرز واژه‌ها و جمله‌ها

کارایی هر سامانه‌ی پردازش متن در تعیین مرز واژه‌ها و جمله‌ها، ارتباط مستقیمی با دقت کل سامانه دارد چرا که نحوه‌ی شناسایی آنها به عنوان عناصر پایه‌ی پردازشی، رابطه‌ی تنگاتنگی با عملکرد پیمانه‌های دیگر سامانه پردازش متن دارد. در بیشتر مواقع، تعیین مرز جمله‌ها از طریق بررسی علایم جداکننده انجام می‌شود. علایمی که برای تعیین مرز جمله از آنها استفاده می‌شود، عبارتند از: ".“، ”؛“، ”!“، ”؟“، ”?“ و ”:“. باید توجه داشت که جستجو برای یافتن این علایم به تنهایی کافی نیست و در بسیاری موارد مشکلاتی را به همراه

دارد [۹]. به عنوان مثال، در زبان فارسی بعضی از واژه‌های اختصاری به صورت چند حرف که با نقطه از هم جدا شده‌اند، ظاهر می‌شوند (مانند «ه.ق.» که مخفف «هجری قمری» است). در صورتی که تعداد این حالات مشکل‌ساز در پیکره زیاد باشد، باید از روش‌های پیچیده‌تری استفاده شود، که قادر به شناسایی این موارد باشند.

شناسایی واژه‌ها نیز از طریق بررسی علایم قابل انجام است. این علایم عبارتند از: فضای خالی، علامت پرش، علامت خط جدید، ”،“، ”,“، ”>“، ”<“، ”]“، ”[“، ”-“، ”_“ و ”/“. در اینجا هم بررسی علایم به تنهایی کافی نیست و معمولا برای بهبود کارایی باید از روش‌های پیچیده‌تری استفاده نمود [۹]. به عنوان مثال، تعیین مرز بعضی فعل‌ها (مانند «جارو زدن»، «می باشد»)، واژه‌های مرکب (مانند «ساده زیست») و واژه‌های به هم چسبیده (مانند «دربرابرباد») به این روش امکان‌پذیر نیست.

۳-۲- یکسان‌سازی متن

یکسان‌سازی نیز به عنوان یکی از اعمال پیش‌پردازشی شناخته می‌شود. گاهی نویسه‌های به کار رفته در دو واژه‌ی یکسان، با هم متفاوت هستند؛ این باعث می‌شود که هنگام شمردن واژه‌ها، دو واژه‌ی یکسان با املاهای متفاوت به عنوان دو واژه‌ی مختلف در نظر گرفته شوند. برای جلوگیری از بروز این مشکل، نیازمند یکدست‌سازی پیکره‌ی متنی هستیم. به عنوان مثال، پیشوند «می» و پسوند «ها» در ابتدا و انتهای واژه‌ها، ممکن است به سه صورت مختلف زیر دیده شوند:

چسبان	جدا با فاصله	جدا بدون فاصله
کتابها	کتاب ها	کتاب‌ها
میرود	می رود	می‌رود

و یا واژه‌ها «مسئول»، «مجموعهٔ» و «پاییز» به صورت‌های زیر در پیکره‌ی متنی دیده شوند:

مسئول ، مسؤول و مسوول

مجموعه‌ی ، مجموعهٔ و مجموعه

پاییز و پائیز

به عنوان یک مثال از یکدست‌سازی، در مورد پسوند «ها» می‌توان شکل‌های «جدا با فاصله» و «جدا بدون فاصله» را نادیده گرفت و با مشاهده‌ی هر یک از این دو، آن را به شکل چسبان در آورد. در [۱۰]، یک مجموعه تبدیل‌های نسبتا جامع و موثر برای یکدست‌سازی متون فارسی ارایه شده است. تجربه نشان می‌دهد که با استفاده از تبدیل‌های زیر، می‌توان پیکره‌ی متنی را به خوبی یکدست‌سازی کرد:

- تبدیل نویسه‌های «ی» و «ک» عربی به نوع فارسی آن.
- تبدیل نویسه‌های «ؤ» به «و»، «ئ» به «ی» و «أ» به «ا».
- تبدیل نویسه‌های «ۀ» و «ۀ» به «ه» در آخر واژه‌ها.
- حذف «ی» از آخر واژه‌هایی مانند «خانه ی».
- حذف فتحه، کسره و ضمه (نویسه‌های ـَ ، ـِ ، ـُ) از واژه‌ها.
- حذف تنوین (نویسه‌های ـً ، ـٍ ، ـٌ) از انتهای واژه‌ها.
- حذف تشدید (شناسه‌ی ـّ) از واژه‌ها.
- حذف شناسه‌ی «ء» در آخر بعضی واژه‌ها مانند «شهداء».
- چسباندن پیشوندهای «می»، «درمی»، «برمی»، «نمی» و «بی» به ابتدای واژه‌ها.
- چسباندن پیشوند «هم» به واژه‌های مانند «هم چنین» به ابتدای واژه‌ها.
- چسباندن پیشوند «به» به واژه‌هایی مانند «به ندرت» به ابتدای واژه‌ها.
- چسباندن پسوندهای «ها»، «های»، «هایی»، «هایم»، «هایت»، «هایش»، «هایمان»، «هایتان» و «هایشان» به انتهای واژه‌ها.
- چسباندن پسوندهای «تر» و «ترین» به آخر واژه‌ها.
- حذف فاصله بعد از پیشوند «بر» در واژه‌هایی مانند «بر می‌گردد».

- تبدیل واژه‌هایی مانند «مسیول» و «مسیله» به «مسوول» و «مساله». واژه‌های «مسئول» و «مسئله»، با تبدیل دو به «مسیول» و «مسیله» تبدیل می‌شوند.
- حذف نویسه‌ی «ـ» که برای کشش نویسه‌های چسبان مورد استفاده قرار می‌گیرد. مانند تبدیل «بـــر» و «بـر» به «بر». معمولاً از این نویسه برای تراز کردن طول خط‌ها استفاده می‌شود.

۳-۳- حذف واژه‌های عمومی

واژه‌های عمومی، واژه‌هایی هستند که پرتکرارند، اما مهم نیستند. در صورتی که فراوانی واژه‌ها را گواهی بر اهمیت آنها بدانیم، ضروری است که واژه‌های عمومی که پرتکرار و بی‌اهمیت هستند را مشخص کنیم. از آنجا که واژه‌های عمومی محدود هستند، پیدا کردن آنها چندان مشکل نیست و تهیه یک لیست برای شناسایی آنها بسنده می‌کند. جدول ۱، تعدادی از این واژه‌ها را نشان می‌دهد که بیشتر از میان حروف اضافه انتخاب شده‌اند. مجموعه واژه‌های عمومی مورد استفاده در پیاده‌سازی شامل بیش از ۵۰۰ واژه است. این مجموعه، واژه‌های عمومی مرسوم هستند که تعدادی از واژه‌های کم ارزش و پرتکرار در پیکره‌ی خبری مانند «خبرگزاری»، «خبرنگار»، «دانشجویان» و «گزارش» به آنها افزوده شده است.

جدول ۱) تعدادی از واژه‌های عمومی

در	نیز	برای	یا	را
به	تا	ها	دو	های
از	ما	آن	آنها	و
که	باید	وی	اما	نمی
این	اند	یک	دیگر	هر
با	هم	خود	اگر	ای

۳-۴- شناسایی عناصر متنی چندتایی

تاکنون تعریف‌های مختلفی برای عناصر متنی چند واژه‌ای ارائه شده است. [۱۰]، یک عنصر متنی چند واژه‌ای را دنباله‌ای از دو یا چند واژه می‌داند که از ویژگی‌های یک عنصر نحوی و معنایی برخوردار است و معنای آن از معنی تک تک عناصرش بدست نمی‌آید. همچنین در [۶] تعریف کاربردی‌تری برای عنصر متنی چندتایی پیشنهاد شده که عبارت است از: دنباله‌ای از واژه‌ها که رخداد همزمان آنها بیش از میزانی است که در کاربردهای عادی و به شکل تصادفی انتظار داریم. به عنوان مثال، «محمدرضا شجریان» از دو بخش تشکیل شده است و با مشاهده‌ی بخش اول آن یعنی «محمدرضا»، احتمال آمدن بخش دوم یعنی «شجریان» بیش از آن مقداری است که انتظار می‌رود به صورت تصادفی پس از بخش اول ظاهر شود.

روش به کار رفته در این مقاله برای شناسایی عناصر متنی چندتایی، استفاده از احتمال شرطی متقارن است. احتمال شرطی متقارن به صورت زیر تعریف می‌شود:

(۱)

$$SPC(w_1,w_2)=\frac{p(w_1,w_2)^2}{p(w_1)\times p(w_2)}$$

که در آن $P(w_1,w_2)$ احتمال رخداد دو واژه‌ی w_1 و w_2 به صورت متوالی است و $P(w_1)$ و $P(w_2)$ نیز به ترتیب احتمال رخداد w_1 و w_2 به تنهایی هستند. از آنجا که هر یک از احتمال‌های نامبرده متناسب با فراوانی واژه در پیکره‌ی متنی هستند، می‌توان به طور مستقیم از فراوانی واژه‌ها استفاده کرد. رابطه‌ی ۱ برای شناسایی عناصر متنی دوتایی استفاده می‌شود، حال آنکه یک عنصر متنی چندتایی می‌تواند شامل بیش از دو واژه باشد. پس از تعیین احتمال شرطی متقارن برای هر مجموعه واژه‌ی کاندید، مقدار بدست آمده با یک آستانه مقایسه می‌شود و چنانچه از آن بیشتر باشد به عنوان یک عنصر متنی چندتایی در نظر گرفته می‌شود. در پیاده‌سازی از آستانه‌های ۰,۰۱ و ۰,۰۰۱ به ترتیب برای شناسایی عناصر متنی دوتایی و سه‌تایی استفاده شد. عناصر متنی دوتایی بدست آمده به صورت خودکار، مورد اصلاح و بازنگری (حذف مواردی که نادرست گزینش شده‌اند) قرار گرفتند و در نهایت لیستی شامل ۱۲۳۱ عنصر حاصل شد. به طور مشابه، لیستی شامل ۷۹۴ عنصر متنی سه‌تایی نیز بدست آمد. جدول ۲ ، تعدادی از عناصر متنی چندتایی شناسایی شده را نشان می‌دهد.

جدول ۲) تعدادی از عناصر متنی دوتایی و سه‌تایی موجود در پیکره‌ی متنی ایسنا

احمدرضا عابدزاده	سیاه پوشان ایتالیا
استادیوم آزادی	سید محمد خاتمی
افریقای جنوبی	شیرجه و واترپلو
ال کلاسیکو	صربستان و مونتنگرو
امام خمینی	فدراسیون فوتبال المان
ایمان مبعلی	لیگ برتر انگلستان
برانکو ایوانکوویچ	محمد مایلی کهن
بوندس لیگا	مهدی مهدوی کیا
حبیب کاشانی	نیروی هوایی عراق
داور وسط	ورزشگاه آزادی تهران

۴- روش پیشنهادی

روش پیشنهادی، یک روش مبتنی بر خوشه‌بندی است. در این روش، ابتدا تمام جمله‌های بدست آمده از همه‌ی متون اولیه خوشه‌بندی می‌شوند. خوشه‌بندی با استفاده از روش خوشه‌بندی k-mean که یک روش خوشه‌بندی ساده است، انجام می‌شود. در نسخه‌ی اصلی روش k-mean تعداد خوشه‌ها از قبل مشخص است و در هر مرحله، خوشه‌ها را با توجه به تعدادی مرکز خوشه تقسیم‌بندی می‌کند و در انتهای هر مرحله مرکز خوشه‌ها را بروز رسانی می‌کند. معیاری شباهتی که در خوشه‌بندی مورد استفاده قرار گرفته، معیار کسینوسی است که مبتنی بر ضرب عناصر متناظر و جمله آنهاست.

پس از خوشه‌بندی، به ازای هر خوشه جمله‌ای که بیشترین ارتباط با سایر جمله‌ها را دارد، گزینش می‌شود. برای این کار، جمله‌ای که حاصل جمع شباهت کسینوسی آن با سایر اعضای خوشه از بقیه بیشتر است به عنوان نماینده‌ی خوشه انتخاب می‌شود.

در آخرین مرحله‌ی خلاصه‌سازی، جمله‌ها با توجه به ترتیب زمانی متن‌ها (خبری) در خلاصه‌ی نهایی درج می‌شوند.

۵- نتایج عملی

روش پیشنهادی روی چند مجموعه متن خبر پیاده‌سازی شد. هر یک از این مجموعه‌ها شامل تعدادی خبر ورزشی هم موضوع است. خبرها از یک مجموعه‌ی بزرگ‌تر شامل ۱۸۰۰۰ خبر ورزشی مربوط به خبرگزاری ایسنا تهیه شده‌اند.

از ۱۰ مجموعه برای ارزیابی سامانه‌ی خلاصه‌سازی پیاده‌سازی استفاده شد. هر مجموعه شامل ۶ تا ۲۳ خبر بود که موضوع بعضی از آنها عبارتند از: «انتقال رونالدینیو از بارسلونا به میلان»، «قهرمانی سپاهان در لیگ برتر»، «تمدید قرارداد مجتبی جباری با استقلال»، «مبارزه با دوپینگ» و ... . شکل ۱، نمونه‌ای از خروجی سامانه‌ی خلاصه‌سازی پیاده‌سازی شده با روش پیشنهادی را نشان می‌دهد.

تاریخ: ۲۰۰۷/۰۸/۰۳ , ساعت: ۶:۵۳
اکنون بستگي دارد نظر برهاني چه باشد.

تاریخ: ۲۰۰۷/۰۸/۰۳ , ساعت: ۱۱:۴۸
مجتبي جباري بازيكن استقلال تهران پس از گذراندن دوران درمان در هلند به كشور بازگشت و در يك روز به دو باشگاه بزرگ تهراني پرسپوليس و استقلال تهران رفت، جباري تصور نمي كرد استقبال از وي در باشگاه استقلال به گونه اي باشد كه مجبور شود قيد حضور در اين تيم را بزند.

تاریخ: ۲۰۰۷/۰۸/۰۴ , ساعت: ۷:۰۱
او بدون اين كه از تيم جديدش نام ببرد، گفت: به احتمال زياد فردا با تيم جديدم قرارداد مي بندم.

تاریخ: ۲۰۰۷/۰۸/۰۴ , ساعت: ۷:۳۲
البته تنها بحث باشگاه پرسپوليس نيست و مذاكراتي نيز با باشگاه سايپا صورت گرفته است.

تاریخ: ۲۰۰۷/۰۸/۰۵ , ساعت: ۵:۵۳
مجتبي جباري رسما با تيم فوتبال استقلال تهران قرارداد يك ساله امضا كرد.

تاریخ: ۲۰۰۷/۰۸/۰۵ , ساعت: ۹:۴۲
وي افزود: پس از اولين جلسه اي كه با جباري برگزار شد و به دليل اختلاف نظرهاي ايجاد شده، مسوولان باشگاه شنبه شب مجددا با جباري صحبت كردند و در نهايت در فضاي كاملا منطقي مقرر شد جباري براي يك سال ديگر در استقلال بازي كند.

شکل ۱) نمونه‌ای از خروجی روش پیشنهادی

در پیاده‌سازی تعداد خوشه‌های ۸ در نظر گرفته شد. همچنین پس از خوشه‌بندی، خوشه‌های که تنها شامل یک جمله بودن (تنها مرکز خوشه) نادیده گرفته شدند. به این ترتیب، تعداد جمله‌های خلاصه‌های پایانی خلاصه‌ساز، ۸ جمله یا کمتر بود (شکل ۱ شامل ۶ جمله است).

برای ارزیابی نتایج از یک روش ذهنی (subjective) که مبتنی بر قضاوت انسانی است، استفاده شد. ارزیابی به این صورت انجام شد که برای خلاصه‌ی هر یک از ۱۰ مجموعه، ۳ داور انسانی رای خود را به صورت خوب، متوسط یا ضعیف اعلام کرد.

خوب	متوسط	ضعیف
۰,۳۳	۰,۵۱	۰,۱۶

۶- نتیجه‌گیری

در این مقاله، یک روش جدید مبتنی بر خوشه‌بندی برای خلاصه‌سازی چندسندی متون فارسی پیشنهاد شد. در این روش، پس از پیش پردازش متن شامل تعیین مرز واژه‌ها و جمله‌ها، یکسان‌سازی متن، حذف واژه‌های عمومی و شناسایی عناصر متنی چندتایی، فرآیند اصلی خلاصه‌سازی آغاز می‌شود. در مرحله‌ی خلاصه‌سازی، ابتدا جمله‌ها خوشه‌بندی می‌شود و سپس به ازای هر خوشه جمله‌ای که بیشترین ارتباط با سایر جمله‌ها را دارد، گزینش می‌شود. در آخرین مرحله‌ی خلاصه‌سازی، جمله‌ها با توجه به ترتیب زمانی متن‌ها (خبری) در خلاصه‌ی نهایی درج می‌شوند. نتایج پیاده‌سازی نشان می‌دهند که در بیشتر موارد خروجی سامانه‌ی خلاصه‌سازی پیشنهادی خلاصه‌ی قابل قبولی را تولید می‌کند (بیش از ۸۰ درصد).

برای بهبود روش پیشنهادی می‌توان از معیارهای دیگری نظیر مکان جمله‌ها، امتیازدهی به جمله‌هایی که عبارات اشاره دارند، امتیازدهی بر حسب به طول جمله، اسم‌های خاص و نقل‌وقول و ... همراه با معیار بیشینه ارتباط درون خوشه‌ای (روش امتیازدهی مورد استفاده در روش پیشنهادی) استفاده کرد.

مراجع

[1] Visser W.T., Wieling M.B., "Sentence-Based Summarization of Scientific Documents", M.S. Project, University of Groningen.

[۲] شمس‌فرد م.، پردازش متون فارسی: دستاوردهای گذشته، چالش‌های پیش رو، دومین کارگاه پژوهشی زبان فارسی و رایانه، ص. ۱۷۲ تا ۱۸۹، تهران ۱۳۸۵.

[3] Mazdak N., Hassel, M.."FarsiSum-a persian text summarizer", Master thesis, Department of linguistics, Stockholm University, 2004.

[4] Dalianis H.," SweSum - A Text Summarizer for Swedish, Technical report", TRITANA-P0015, IPLab-174, NADA, KTH, October 2000.

[۵] کریمی ز.، شمس‌فرد م.، «سیستم خلاصه‌سازی خودکار متون فارسی»، دوازدهمین کنفرانس بین‌المللی انجمن کامپیوتر ایران، تهران ۱۳۸۵.

[۶] شهابی ا.ش.، «چکیده‌سازی متون زبان فارسی»، دومین کنفرانس علوم شناختی، صفحه ۵۶، تهران ۱۳۸۱.

[۷] ورفریاری و.، «بررسی جامع سیستم‌های خلاصه‌ساز فارسی و تولید یک نمونه عملی»، پایان‌نامه کارشناسی ارشد، دانشکده‌ی مهندسی کامپیوتر، دانشگاه علم و صنعت ایران ۱۳۷۶.

[۸] مشکی م.، «بررسی روش‌های خلاصه‌سازی متون غیرساخت‌یافته‌ی فارسی»، سمینار کارشناسی ارشد، دانشکده‌ی مهندسی کامپیوتر، دانشگاه علم و صنعت ایران ۱۳۸۶.

[۹] ایران‌پور مبارکه م.، «بررسی مشکلات تعیین حدود جمله و کلمه»، سمینار کارشناسی ارشد، دانشگاه علم و صنعت ایران، ۱۳۸۶.

[۱۰] حافظی م.م. ، ثامتی ح.، منصوری ن.، منتظری ن.، بحرانی م.، موثق ح.، «ارائه یک مدل دستوری برای بهبود دقت سیستم‌های بازشناسی گفتار پیوسته‌ی فارسی»، دومین کارگاه پژوهشی زبان فارسی و رایانه، ص. ۸۰ تا ۹۱، تهران ۱۳۸۵.

[11] Choueka, Y. (1988). Looking for needles in a haystack, or locating interesting collocational expressions in large textual databases. In roceedings of RIAO'88 Conference, Cambridge, MA, pp. 609-623.

کارگاه‌های آموزشی مرکز اطلاعات علمی

آموزش مهارت های کاربردی در تدوین و چاپ مقاله

اصول تنظیم قراردادها

مقاله نویسی علوم انسانی