خلاصه سازی خودکار چند سندی مبتنی بر مفاهیم

آصف پورمعصومی، محسن کاهانی، محسن کامیار، حسین کامیار

as.poormasoomi@stu-mail.um.ac.ir, kahani@um.ac.ir, {mohsen.kamyar, hossein.kamyar}@stu-mail.um.ac.ir

آزمایشگاه فناوری وب، دانشگاه فردوسی مشهد

چکیده

خلاصه سازی خودکار چند سندی، روشی برای ارائه فشرده مطالبی است که در ارتباط با یک موضوع بوده ولی جهت دید آنها متفاوت از یکدیگر می باشند. خلاصه خوب، باید بیانگر زمینه کلی بوده و ضمن بیان دیدگاههای مختلف موجود در متن از خوانایی و پیوستگی بالایی برخوردار باشد. در این مقاله با بهره‌گیری شکل جدیدی از روش استخراج روابط معنایی موجود در متن (LSA یا LSI) و تکنیک برچسب زنی معنایی نقش لغات(SRL)، روشی جدید برای خلاصه‌سازی چند سندی شده است. در ابتدا با استفاده از ماتریس کلمه- سند به جای ماتریس کلمه- جمله و با بهره گیری از LSA، جملات مهم مرتبط با زمینه استخراج می شود. در گام بعدی با استفاده از تکنیک SRL و با استفاده از WordNet شباهت معنایی جملات استخراج شده و در نهایت جملات شبیه به هم حذف می شوند. نتایج آزمایش‌ها برروی داده‌های DUC2007 حاکی از افزایش چشمگیر دقت در قیاس با روش‌های قبلی مبتنی بر LSA و سیستم‌های موجود در DUC2007 می باشد.

کلمات کلیدی : خلاصه سازی چند سندی، LSA، SRL

۱- مقدمه

خلاصه‌سازی خودکار چند سندی به فرآیند تولید یک خلاصه فشرده از اسناد با حفظ موضوعیت، خوانایی و پیوستگی مطالب اتلاق می شود.[1] افزایش قارچ گونه اسناد الکترونیکی در موضوعات شبیه به یکدیگر و مشکلات کمبود زمان برای عموم کاربران، منجر به استقبال روز افزون دانش پژوهان جهت تحقیق در زمینه خلاصه‌سازی خودکار چندسندی شده است. خلاصه‌سازی چندسندی، ارتباط تنگاتنگی با مباحث سیستم‌های پاسخگو[۱] و خلاصه‌سازی مبتنی بر پرس‌وجو دارد.[2] در حقیقت خلاصه‌سازی چندسندی برروی اسنادی انجام می شود که در ارتباط با یک موضوع هستند ولی جهت دید آنها متفاوت از یکدیگر است. به عنوان مثال موضوع "*مشکل جهانی کمبود آب*" را در نظر بگیرید. در ارتباط با این موضوع اسناد مختلفی می توانیم داشته باشیم که از دیدگاههای متفاوت به این موضوع پرداخته باشند. مثلا یکی در مورد "کمبود آب در ایران" و دیگری در خصوص "کمبود آب در پاکستان" باشد.

در خلاصه‌سازی چندسندی با پیچیدگی‌های بیشتری نسبت به خلاصه سازی تک‌سندی رویرو هستیم. مهمترین چالش‌های پیش رو در این دسته از خلاصه سازها عبارتند از [3]:

- چون اسناد با دیدگاههای متفاوت به شرح یک موضوع می پردازند که گاها متناقض با یکدیگر هم بوده، بنابراین تولید خلاصه‌ای با خوانایی بالا امری دشوار خواهد بود.
- با توجه به این که همه اسناد در ارتباط با یک موضوع کلی می باشند، بحث همپوشانی جمله‌های انتخاب شده از این اسناد یا همان افزونگی[۲] مشکل مهمی می باشد.
- استخراج دیدگاههای مختلف پنهان در اسناد مختلف و پوشش دادن تمامی آنها نیاز به دقت و توجه فراوان دارد.

به صورت کلی روش‌های خلاصه‌سازی را می‌توان به دو دسته نظارت محور[۳] و غیرنظارت[۴]شونده تقسیم نمود[1]. در روش‌های نظارت محور به منظور رسیدن به دقت کافی، نیاز به مجموعه بزرگی از خلاصه‌های تولید شده توسط انسان می‌باشد. این دسته از خلاصه‌ها مدل‌گرا بوده و طبیعتا با تغییر نوع داده‌ها و ویژگی‌های آنها، نیاز به تولید مجدد داده-های آموزشی دارند[4][5]. روش‌های غیرنظارت شونده نیازی به خلاصه‌های انسانی برای آموزش ندارند[6].

در سال‌های اخیر مقالات زیادی از LSA[9][۵] برای خلاصه‌سازی استفاده کرده‌اند [6]. LSA برای حل مشکل تنک[۶] بودن داده‌ها ارائه شده است. این روش با نمایش داده‌ها در فضای معنایی کوچکتر و در حقیقت کاهش ابعاد، تا حد زیادی مشکل داده‌های تنک را حل می کند. این کاهش ابعاد منجر به بهبود کارایی در بسیاری از کاربردها شده است.[6][9] در تمامی روش‌هایی که از LSA برای خلاصه سازی استفاده می کنند بردار جملات را بااستفاده از اعمال SVD[۷] برروی ماتریس *کلمه- جمله* می‌سازند. LSA جملات را مستقیما با استفاده از این ماتریس و با توجه به ویژگی‌های معنایی آنها مرتب و دسته‌بندی می کند.

بر خلاف روش‌های قبلی، در روش جدیدی که در این مقاله ارائه شده به جای استفاده از ماتریس *کلمه–جمله* و سپس انتخاب مستقیم جملات توسط LSA، از ماتریس *کلمه– سند* استفاده شده و در طی دو مرحله جملات مهم استخراج می‌گردد. در گام نخست مفاهیم اصلی موجود در پیکره اسناد با استفاده از روش LSA و ماتریس کلمه- سند استخراج می‌شود. درگام بعدی این مفاهیم به دسته‌های موضوعی

انتساب داده می‌شوند. سپس جملات بر اساس میزان شباهتشان با مفاهیم موجود مرتب می‌شوند. در انتها هم با استفاده از تکنیک SRL[8] و با محاسبه شباهت معنایی جملات، جملات تکراری حذف می‌گردند. نتایج حاصل از آزمایش‌های انجام شده بر داده‌های DUC2007 نشان گر بهبود قابل توجه نسبت به روش‌های پیشین می‌باشد.

ساختار این مقاله به شرح زیر هست. در فصل بعد مروری بر ادبیات موضوع شده است. در فصل ۳روش پیشنهادی شرح داده شده و در فصل ۴ نتایج ارزیابی سیستم‌های قبلی آورده شده است. در انتهای مقاله جمع بندی و کارهای پیش رو توضیح داده شده است.

۲- مرور ادبیات

در این بخش به صورت مختصر به روش‌های مبتنی بر LSA در خلاصه‌سازی چندسندی اشاره شده است. اولین بار Gong در سال ۲۰۰۱ از LSA در خلاصه‌سازی استفاده نمود[6] وی در ابتدا ماتریس کلمه-*جمله* را با محاسبه معیار $TFISF$ تشکیل داد. $TFISF$ معیار محاسبه وزن در واحد *جمله* بوده و به صورت زیر محاسبه می شود

$$(TF-ISF)_{i,j} = \left(\frac{tf_{i,j}}{\sum_k tf_{k,j}}\right) Log_2\left(\frac{|S|}{|\{s: t_i \in s\}|}\right)$$

$tf_{i,j}$ فرکانس نسبی کلمه i در جمله j و $|S|$ تعداد جملات موجود در پیکره می باشد. پس از محاسبه معیار $TFISF$ ماتریس کلمه-جمله تشکیل داده شده و سپس با بهره‌گیری از LSA مهمترین جملات استخراج می شود. مشکل اصلی این روش این است که معیارهای وزن دهی نظیر $TFISF$ قادر به بیان مفاهیم اصلی پنهان از دیدگاه کلی نمی باشد چراکه در معیار وزن دهی $TFISF$، فرکانس کلمات در سطح جمله مورد بررسی قرار می گیرد و تاثیر جملات بر یکدیگر در سطوح بالاتر لحاظ نمی شود.

Steinberger [7] با دخیل کردن روش ادغام ضمایر (anaphora resolution) در LSA، کارایی روش Gong را افزایش داد. وی در این مقاله ثابت کرد که ادغام ضمایر با استفاده از روش افزایشی، دقت خلاصه‌سازی روش Gong را افزایش می دهد.

Yeh [5]در سال ۲۰۰۵، با استفاده توامان از LSA و T.R.M(نگاشت رابطه ای متن) ساختار معنایی پنهان موجود در سند را استخراج کرده و خلاصه را بر اساس این روابط پنهانی استخراج شده تولید نمود.

Yu [8] در سال ۲۰۰۹ روشی برای خلاصه‌سازی اخبار مبتنی بر LSA ارائه نمود. وی در ابتدا با استفاده از LSA شباهت معنایی بین جملات را محاسبه کرده و سپس با استفاده از روش خوشه بندی مبتنی بر روش k-means جملات را دسته‌بندی کرده و سپس از هر کلاستر، جمله‌ای که بیشترین شباهت را به موضوع داشته باشد به عنوان نماینده آن خوشه بر می‌گرداند.

در تمامی این روش‌ها با محوریت قرار دادن جمله و ساخت ماتریس کلمه- *جمله*، از LSA استفاده شده است. در روش پیشنهاد شده در این مقاله با محوریت قرار دادن *سند* به عنوان عنصر کلیدی برای استخراج زمینه، خلاصه را در دو مرحله تولید می کنیم.

۳- روش پیشنهادی

۳-۱- فازهای کلی روش ارائه شده

الف) *فاز پیش پردازش :* در این فاز عملیات زیر باید انجام گیرد

- در ابتدا جملات استخراج شده و پس از حذف کلمات پرتکرار بی ارزش[9]، عمل ریشه‌یابی انجام می شود. الگوریتم-های ریشه‌یابی زیادی تاکنون ارائه شده که معروفترین آنها الگوریتم پرتر[10] می باشد.
- پس از استخراج کلمات، ماتریس کلمه-سند ساخته می‌شود. از معیارهای وزن‌دهی مختلفی در این مرحله می‌توان استفاده کرد. در این مقاله از معیار وزن‌دهی $TFIDF$ که به صورت زیر محاسبه می شود استفاده شده است

$$(TF-IDF)_{i,j} = \left(\frac{tf_{i,j}}{\sum_k tf_{k,j}}\right) Log_2\left(\frac{|D|}{|\{d: t_i \in d\}|}\right)$$

$tf_{i,j}$ تعداد دفعات تکرار کلمه i در سند j و $|D|$ کل اسناد موجود در پیکره می‌باشد.

ب) *فاز استخراج مفاهیم اصلی موجود در اسناد:* در این مرحله با استفاده از LSA، مفاهیم اصلی موجود در کل پیکره استخراج می-گردد. در گام بعدی با اندازه‌گیری فاصله کسینوسی بردار مفاهیم و بردار اسناد، میزان شباهت هر مفهوم[11] به هر موضوع محاسبه شده و بدین ترتیب مفهوم اصلی یا همان زمینه[12] یک موضوع بدست می‌آید. سپس مهمترین جملات مرتبط با زمینه اصلی استخراج می‌شود. به همین منظور شباهت بین بردار فرکانس کلمه‌های جملات و بردار مفهوم منتسب داده شده به موضوع محاسبه شده و جمله‌ها بر اساس میزان شباهتشان به زمینه اصلی متن، به صورت نزولی مرتب می‌شوند.

ج) *محاسبه میزان شباهت معنایی برای کشف جملات مشابه* : در این مرحله با استفاده از برچسب‌زنی معنایی نقش کلمات، نقش واحدهای معنایی جمله استخراج شده و سپس با اندازه‌گیری شباهت معنایی کلمات موجود در نقش‌های بدست آمده با استفاده از شبکه واژگان[10]، شباهت بین جمله‌ها محاسبه می شود.

د) *تولید خلاصه :* درنهایت با تشخیص جملاتی که از لحاظ معنا شبیه به هم هستند و با در نظر گرفتن میزان فشرده سازی، جملات تکراری حذف شده و بدین ترتیب دیدگاههای مختلف پوشش داده می شود.

۳-۲- استخراج مفاهیم اصلی موجود در اسناد

LSA در ابتدا برای حل مشکل هم خانواده‌ها و هم آوایی‌ها[13] در IR معرفی شد[9]. از آن پس LSA در کانون توجه محققان در زمینه پردازش متن قرار گرفت و در بسیاری از مقالات به صورت تئوری و

عملی آنالیز گردید[11]. LSA در درون خود از روش SVD[14] استفاده می‌کند. SVD ماتریس A با ابعاد m×n را به سه ماتریس $A = USV^T$ تجزیه می‌کند که $U = [u_{ij}]$ ماتریس ارتونرمال[15] ستونی m×m بوده و ستون‌های آن بردارهای یکه سمت چپ نامیده می شوند؛ $S = diagonal(\sigma_1, \sigma_2, \dots, \sigma_n)$ ماتریس قطری n×n است که عناصرقطر اصلی آن مقادیر ویژه غیرمنفی می‌باشند که به صورت نزولی مرتبط شده‌اند و $V = [v_{ij}]$ هم ماتریس ارتونرمال سطری بوده و ستون‌های آن بردارهای یکه سمت راست نامیده می‌شوند. اگر rank(A)= r باشد آنگاه حالت زیر برقرار خواهد بود :

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq \sigma_{r+1} = \dots = \sigma_n = 0$$

از دیدگاه ریاضی، استفاده از SVD برروی یک ماتریس منجر به نگاشت فضای m بعدی به یک فضای r بعدی می شود و در بسیاری از کاربردها از آن برای کاهش ابعاد در ماتریس‌های تنک استفاده می شود. از دیدگاه پردازش زبان طبیعی، SVD باعث استخراج روابط معنایی پنهان در ساختار داده می‌شود. اولین بار Gong در سال ۲۰۰۱ از LSA در خلاصه‌سازی استفاده نمود[6]. در روش Gong و سایر روش‌های اشاره شده، بر ویژگی‌های جمله به عنوان عنصر محوری تمرکز شده است. در این روش‌ها، با تولید بردار فرکانس نسبی جملات از طریق محاسبه معیارهای وزن‌دهی در *واحد جمله* نظیر *TFISF*، ماتریس کلمه-جمله تولید می‌شود. سپس SVD برروی این ماتریس اعمال شده و با استخراج ماتریس V که می‌توان آن را ماتریس مفهوم-جمله نامید، بیشترین مقادیر موجود در هر ستون (مفهوم) به عنوان بهترین جمله‌ای که نشانگر آن مفهوم است، انتخاب و در خلاصه قرار داده می‌شود. ایراد تمامی این روش‌ها عدم توجه به مفاهیم اصلی پنهان در *کل سند* می‌باشد و این در حالی است که یک خلاصه خوب در خلاصه‌سازی چندسندی، باید ضمن بیان موضوع و زمینه اصلی موجود در اسناد، دیدگاههای مختلف مرتبط با زمینه را نشان دهد.

در این مقاله با ارائه روشی جدید سعی شده تا این مشکلات برطرف شوند. به همین منظور در ابتدا به جای استفاده از *TFISF*، وزن *TFIDF* برای کلمات در *واحد سند* محاسبه شده و سپس ماتریس کلمه– سند ایجاد می‌شود. در گام بعدی SVD برروی این ماتریس اعمال شده و ماتریس بردارهای ویژه استخراج می‌شود. در مرحله بعد، از ماتریس ستونی U استفاده می‌شود. در ماتریس U، بردارهای ستونی مستقل خطی هستند. این بردارهای ستونی از دید پردازش زبان طبیعی همان مفاهیم مستقل از هم می‌باشند. بنابراین از دید معنایی، ماتریس U، ماتریس کلمه- مفهوم است. حال باید این مفاهیم به کلاسترهای موضوعی انتساب داده شوند. برای اینکار فاصله کسینوسی بین بردارهای مفاهیم $C_i = [c_{i1}, c_{i2}, \dots, c_{im}]$ و بردارهای اسناد $D_j = [d_{j1}, d_{j2}, \dots, d_{jm}]$ محاسبه می‌شود. فاصله کسینوسی بین این دو بردار به صورت زیر محاسبه می شود :

$$cos(C_i, D_j) = \frac{\sum_k c_{ik} d_{jk}}{\sqrt{\sum_k (c_{ik})^2} \times \sqrt{\sum_k (d_{jk})^2}}$$

پس از انتساب مفاهیم به موضوعات موجود در پیکره، فاصله بین جملات سندهای یک موضوع با مفهوم انتساب داده شده به آن موضوع محاسبه می‌شود. برای این منظور در ابتدا بردار جملات ایجاد شده و سپس مجددا از فاصله کسینوسی بین بردار جملات $S_j = [s_{j1}, s_{j2}, \dots, s_{jm}]$ و بردار مفهوم $C_i = [c_{i1}, c_{i2}, \dots, c_{im}]$ برای مرتب کردن جملات استفاده می‌شود.

۳-۳- اندازه گیری شباهت معنایی

در فاز قبلی جملات بر اساس میزان ارتباطشان با زمینه متن یا همان مفهوم اصلی مرتب شدند. طبیعتا بدلیل ماهیت چندسندی بودن پیکره، بسیاری از این جمله‌ها شبیه به هم هستند. در این فاز باید جملات تکراری از لحاظ معنا حذف شده و دیدگاه‌های مختلف استخراج شوند. در این فاز برخلاف فاز قبلی، استفاده از فاصله کسینوسی نمی تواند به خوبی شباهت بین جملات را نشان دهد. برای روشن‌تر شدن موضوع، جملات زیر را در نظر بگیرید :

S1 = United States Army, successfully tested an anti-missile defense system.

S2 = U.S. military projectile interceptor, streaked into space and hit the target.

S3 = Iran's weekend test of a long-range missile underscored the need for a U.S. national missile defense system.

جملات *S1* و *S2* تقریبا بیانگر یک خبر هستند ولی چون از لحاظ نحوی کلمه مشترکی ندارند فاصله کسینوسی آنها صفر می‌شود (از دید فاصله کسینوسی شبیه هم نیستند): اما با احتساب فاصله کسینوسی، جملات S1 و S3 چون دارای کلمات مشترک هستند، شبیه به هم می‌باشند که البته این تشخیص اشتباه است. بنابراین فاصله کسینوسی بدلیل توجه به شباهت نحوی و عدم توجه به موقعیت و نقش کلمات در جمله، نمی‌تواند به خوبی شباهت بین جملات را محاسبه نماید. استفاده از فاصله کسینوسی در فاز قبل این مشکل را نداشت. چراکه در فاز قبلی، فاصله کسینوسی بین بردار جملات و بردار مفاهیم (که در برگیرنده همه کلمات موجود در متن با وزن خاص خود هست) محاسبه شد و هدف صرفا استخراج جملات مرتبط با زمینه متن بود. برای حل این مشکل، شباهت معنایی بین جملات محاسبه می‌شود.

نقش معنایی نقشی است که یک جز در ارتباط با فعل جمله ایفا می کند. تاکنون چندین ابزار برای برچسب زنی معنایی ارائه شده است. در این مرحله از ابزار SRL دانشگاه Illinois At Urbana Champaign[16] استفاده شده است. با استفاده از این ابزار، نقش‌های معنایی جمله نظیر فاعل، مفعول مستقیم و غیر مستقیم و ... در ارتباط با فعل جمله مشخص می‌شود.

۳-۳-۱- شباهت معنایی مبتنی بر شبکه واژگان جملات

برای اندازه‌گیری شباهت معنایی بین دو جمله، ابتدا اجزای آنها با استفاده از SRL جدا شده و سپس با استفاده از WordNet، ارتباط بین کلمات در نقش‌های معنایی مشخص محاسبه می‌شود. اگر دو کلمه در یک نقش معنایی مشخص، یکسان بوده و یا دارای ارتباطات معنایی نظیر مترادف[۱۷]، شمول[۱۸]، مشمول[۱۹]، جزئی از کل[۲۰]، کل مربوط به جزء[۲۱] باشند، آنگاه از لحاظ معنایی مرتبط با هم خواهند بود. اگر P_{ak} و P_{bl} به ترتیب دو جزء گزاره‌ای از جملات S_a و S_b بوده و $R = \{r_1, r_2, \dots, r_{roleNum}\}$ را مجموعه نقش‌های موجود و $Term_k(r_i)$ را مجموعه کلمات جزء P_k در نقش معنایی r_i در نظر بگیریم آنگاه شباهت معنایی بین دو جمله به صورت زیر محاسبه می‌شود:

$$sim(S_a, S_b) = \frac{\sum_{k=1}^{m}\sum_{l=1}^{n} Semantic_{Sim}(P_{ak}, P_{bl})}{m+n}$$

$$Semantic_{Sim}(P_k, P_l) = \frac{\sum_{i=1}^{roleNum} RelatedT(Term_k(r_i), Term_l(r_i))}{|Term_k(r_i)| + |Term_l(r_i)|}$$

تابع $RelatedT(Term_k(r_i), Term_l(r_i))$، تعداد کلماتی که در دو جزء P_k و P_l از لحاظ معنایی به هم مرتبط اند مشخص می‌کند. میزان شباهت جملات مقداری بین ۰ و ۱ خواهد بود.

۳-۴- تولید خلاصه

الگوریتم زیر برای تولید خلاصه ارائه شده است:

maxLen حداکثر میزان فشرده‌سازی (برحسب کلمه) و list هم لیست جملات مرتب شده در فاز ۲ می‌باشد. sumSent هم جملات خلاصه بوده که در ابتدای اجرای الگوریتم تهی است. شبه کد این الگوریتم در ذیل آورده شده است:

۱. مقادیر k و sLen را به ترتیب برابر ۱ و ۰ قرار بده.
۲. جمله list[k] را انتخاب و با جملات sumSent مقایسه کن و در صورتی که شباهت معنایی آن با تمامی جملات sumSent کمتر از α بود، آن را به sumSent اضافه کن (میزان α به صورت تجربی 0.7 قرار داده شده است. بدلیل کمبود جا از ذکر نمودار تغییرات مربوطه معذوریم).
۳. در صورتیکه sLen<maxLen آنگاه عملیات زیر را انجام بده
   ۱،۳. مقدار k را یک واحد اضافه کن.
   ۲،۳. طول list[k] را به sLen اضافه کن.
   ۳،۳. برو به مرحله ۲.
۴. summSent را به عنوان خلاصه بر گردان.

۴- ارزیابی کارایی

۴-۱- مجموعه داده ها

برای ارزیابی از مجموعه داده‌های DUC2007 استفاده شده است. این مجموعه از داده‌ها مخصوص خلاصه‌سازی چند سندی بوده و شامل ۱۱۲۵ سند می‌باشد. مشخصات کلی این مجموعه داده‌ها را در جدول(۱) مشاهده می‌کنید. این مجموعه داده‌ها توسط Document Understanding Conference و با صرف بیش از ۳۰۰۰ ساعت زمان تهیه شده است

سیستم های پیاده سازی شده :

- LSA : سیستم ارائه شده توسط Gong [6] که مبتنی بر آنالیز کلمه- سند است.
- سیستم پیشنهادی در مقاله با اعمال فاز ۱و۲ به نام our1
- سیستم پیشنهادی در مقاله با اعمال فاز ۱و۲ به نام our2

جدول (۱) – مشخصات کلی مجموعه داده های DUC2007

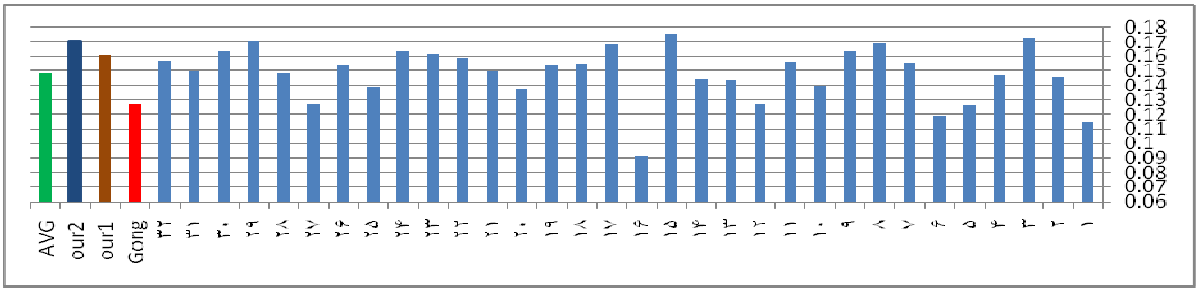
Number of Topics	۴۵
Number of Documents per Topics(Clusters)	۲۵
Number of Terms	۵۳۱۱۷۴
Number of Terms without Stopwords & Stemmings	۲۰۰۵۷
Number of Summarizer Systems	۳۲
Evaluation methods	ROUGE2-ROUGESU4

۴-۲- ابزار ارزیابی

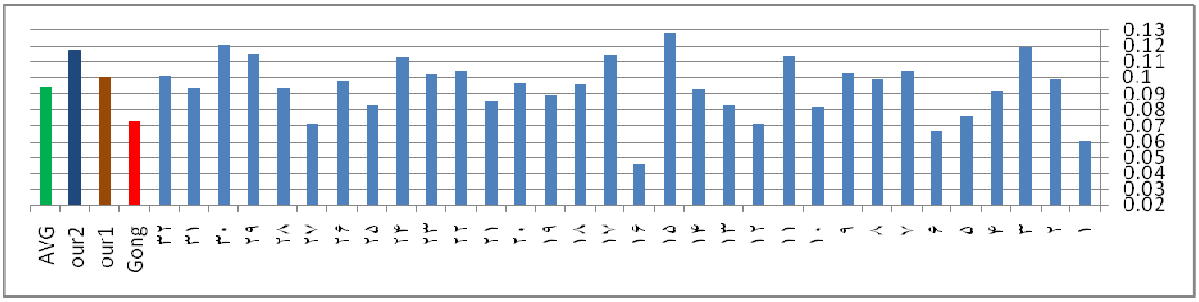
برای ارزیابی از ابزار ROUGE[12][۲۲] استفاده نمودیم. در سال‌های اخیر، در مقالات مختلفی از این ابزار برای ارزیابی استفاده شده است. سیستم‌های موجود در DUC هم با این ابزار ارزیابی شده‌اند. این ابزار شامل معیارهایی برای تعیین کیفیت خلاصه‌ها به صورت خودکار، از طریق مقایسه با خلاصه‌های تولید شده توسط انسان (خلاصه‌های ایده آل) می باشد. از جمله این معیارها به ROUGE-N، ROUGE-L، ROUGE-W و ROUGE-S می‌توان اشاره کرد. با توجه به اینکه در ارزیابی DUC2007 از ROUGE-2 و ROUGE-SU4 برای ارزیابی میانگین سیستم‌ها استفاده شده است بنابراین ما هم از این دو معیار برای ارزیابی استفاده می‌کنیم.

۴-۳- روش ارزیابی

به منظور ارزیابی در ابتدا عملیات پیش‌پردازش را برروی کل پیکره اعمال کرده و پس از تشکیل ماتریس کلمه- سند و استخراج مفاهیم توسط LSA، آنها را به کلاسترها انتساب می‌دهیم. سپس سه کلاستر را به صورت تصادفی از میان ۴۵ تا موضوع انتخاب کرده و سایر عملیات را دنبال می‌کنیم. اشکال(۱) و (۲) نتایج مقایسه سیستم پیشنهاد شده با سایر سیستم‌ها را نشان می‌دهد. همانطور که از نتایج مشخص است پیشرفت قابل توجهی نسبت به روش Gong مشاهده می‌شود. در شکل(۱) سیستم‌ها با معیار Recall ابزار ROUGE-SU4 ارزیابی شده‌اند. ابزار ROUGE مبتنی بر معیار Recall هست و اگر چه که قادر به محاسبه Precision هم هست ولی در ارزیابی‌ها از معیار



شکل(۱) – نتایج ارزیابی معیار Recall برروی داده های DUC2007 با ابزار ROUGE-SU4



شکل(۲) – نتایج ارزیابی معیار Recalll برروی داده های DUC2007 با ابزار ROUGE-2

Recall برای مقایسه سیستم‌ها استفاده می‌کند. در شکل۱و۲، our1 سیستم پیشنهادی با فاز ۱،۲ و our2 با در نظر گرفتن تمامی فازها می‌باشد. AVG هم میانگین Recall سیستم‌های DUC2007 می باشد. در مقالات عموما سیستم‌های پیشنهادی با *میانگین* سیستم‌های DUC که بسیار قدرتمند می‌باشند مقایسه می‌شوند. میزان پیشرفت Recall سیستم نهایی نسبت به روش Gong و میانگین سیستم‌های DUC2007 قابل توجه می‌باشد. لازم به ذکر است که سیستم پیشنهادی با معیار F[۲۳] که ترکیبی از Recall و Precision هست، هم ارزیابی شده که نشان دهنده عملکرد خوب سیستم پیشنهادی است ولی بدلیل کمبود جا این نمودار نشان داده نشده است..

۵- نتیجه گیری و کارهای آتی

در این مقاله با شناخت صحیح ویژگی‌های خلاصه‌سازی چند سندی و با بکارگیری LSA بر روی ماتریس کلمه- سند، زمینه اصلی موضوع استخراج شده و جملات بر اساس شباهتشان با زمینه مرتب شدند. سپس با استفاده از شباهت معنایی جملات تکراری حذف شده و در نهایت خلاصه متناسب با میزان فشرده‌سازی تولید شد. نتایج آزمایشات بیانگر صحیح بودن مسیر کلی طی شده در روند اجرای مقاله می باشد. در آینده در نظر داریم با دخیل کردن پارامترهای دیگری نظیر طول جملات، میزان مقادیر ویژه متناظر با مفاهیم استخراج شده و همچنین در نظر گرفتن موضوع، ضمن افزایش دقت، سیستم خلاصه‌سازی مبتنی بر پرس‌وجو ارائه نماییم.

مراجع

[1] I. Mani. *Automatic summarization*. John Benjamins Publishing Company, 2001.

[2] T. Hirao, Y. Sasaki, and H. Isozaki. An extrinsic evaluation for question-biased text summarization on qa tasks. In *Proceedings of NAACL workshop on Automatic summarization*, 2001.

[3] X. Wan, J. Yang, and J. Xiao. Manifold-ranking based topic-focused multi document summarization. In *Proceedings of IJCAI,* 2007.

[4] Amini, M. R., & Gallinari, P. the use of unlabeled data to improve supervised learning for text summarization. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* , SIGIR`02, 2002.

[5] Yeh, J. Y., Ke, H. R., Yang, W. P., & Meng, I. H. Text summarization using a trainable summarizer and latent semantic analysis. *Information Processing and Management, 41, 75-95,* 2005.

[6] Gong, Y., & Liu, X. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval , SIGIR`01,* New Orleans, 2001.

[7] Steinberger, J., & Kabadjov, M.A. & Poesio, M., & Sanchez-Graillet,O . Improving LSA-based summarization with anaphora resolution. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. 2005.

[8] Yu, H. News summarization based on semantic similarity measure. *Ninth International Conference on Hybrid Intelligent Systems, vol. 1, pp.180-183,* 2009.

[9] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. *Indexing by latent semantic analysis*. Journal of the American Society for Information Science and Technology(JASIS), 41(6):391-470, 1990.

[10] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[11] C. H. Papadimitrious, P. Raghavan, H. Tamaki, and S. Vempala. Latent semantic indexing:A probabilistic analysis. *J. Comput. Syst. Sci., 61(2):217-235, 2000.*

[12] C. –Y. Lin and Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proccedings of NLT-NAACL*, 2003.

---

[1] - question-answering
[2] - redundancy
[3] - supervised
[4] - unsupervised
[5] - latent semantic analysiss
[6] - sparse
[7] - singular value decomposition
[8] - semantic role labeling
[9] - stop words
[10] - porter
[11] - concept
[12] - context
[13] - polysemy
[14] - singular value decomposition
[15] - orthonormal
[16] - http://cogcomp.cs.illinois.edu/
[17] - synonym
[18] - hypernym
[19] - hyponym
[20] - meronym
[21] - holonym
[22] - recall-oriented understudy for gisting evaluation
[23] - $F = ((1+ \beta^2)RP)/(R+ \beta^2 P)$