دوفصلنامه مدیریت مهندسی و رایانش نرم/دوره ۶، شماره ۲، پاییز و زمستان ۱۳۹۹، شماره پیاپی۱۲
صفحات ۱۹۸- ۱۸۳ (DOI): 10.22091/jemsc.2018.1277
JEMSC.QOM
مدیریت مهندسی و رایانش نرم ۱۳۹۴

# خلاصه‌سازی چندسندی استخراجی مبتنی بر پرس‌وجوی متن با استفاده از تفسیر و استلزام متنی*

**علی ناصراسدی**[۱]

## چکیده

*جستجو و اطلاع از محتوای اسناد متنی که گسترده‌ترین نوع اطلاعات بر روی چنین شبکه‌هایی هستند، بسیار مشکل و گاهی اوقات غیرممکن می‌باشد. هدف سیستم‌های خلاصه‌سازی چند سندی متن، تولید کردن خلاصه‌ای با طول ثابت از اسناد متنی ورودی ضمن پوشش حداکثری محتوای اسناد می‌باشد. مقاله حاضر، روشی جدید برای خلاصه‌سازی اسناد متنی بر مبنای استفاده از روابط تفسیر و استلزام متنی و با فرموله‌سازی مسئله در قالب یک مسئله بهینه‌سازی ارائه کرده است. در این روش، جمله‌های درون اسناد ورودی ابتدا بر اساس رابطه تفسیر متنی خوشه‌بندی شده سپس امتیاز استلزام متنی برای کسری از سرآیند خوشه‌ها که دارای بیشترین امتیاز مرتبط با پرس‌وجوی کاربر هستند محاسبه شده و بر اساس آن، امتیاز نهایی هر جمله به دست می‌آید. در نهایت، به کمک دو رویکرد حریصانه و برنامه‌ریزی پویا مسئله بهینه‌سازی حل شده و ضمن انتخاب بهترین جمله‌ها، خلاصه نهایی تولید می‌شود. نتایج اجرای سیستم پیشنهادی بر روی مجموعه‌داده‌های استاندارد و انجام ارزیابی بر اساس سیستم ROUGE نشان می‌دهند که این سیستم کارایی بهترین سیستم‌های خلاصه‌سازی استخراجی مبتنی بر پرس‌وجو را به صورت میانگین حداقل به میزان ۲٪/۵ بهبود داده است.*

***کلمات کلیدی:*** *پردازش زبان طبیعی، خلاصه‌سازی متن، تفسیر متنی، استلزام متنی، کوله‌پشتی صفر و یک.*

---

* تاریخ دریافت: ۹۷/۶/۳۰؛ تاریخ پذیرش: ۹۷/۱۰/۸.

۱. استادیار گروه کامپیوتر، مجتمع آموزش عالی زرند، کرمان، ایران naserasadi@uk.ac.ir

## مقدمه

یکی از پدیده‌های عصر دیجیتالی، وجود حجم زیاد اطلاعات برخط در اشکال مختلف بر روی اینترنت و سایر شبکه‌های کامپیوتری و نرخ روزافزون رشد آنها می‌باشد. در این بین، اسناد متنی از پرکاربردترین اشکال اطلاعات در چنین شبکه‌هایی می‌باشد. یکی از مشکلات ناشی از این پدیده، دسترسی به اطلاعات موثر و در عین حال مختصر برای کاربران شبکه‌های ارتباطی است؛ به گونه‌ای که، با جستجوی یک موضوع خبری در موتورهای جستجو هزاران و شاید میلیون‌ها صفحه وب شامل اطلاعات متنی مرتبط با آن موضوع در اختیار کاربر قرار می‌گیرد. به این ترتیب، مهم‌ترین مشکل کاربر در استفاده از این داده‌ها و اطلاعات، طولانی بودن (و احتمالاً غیرممکن بودن) مطالعه همه این اسناد به منظور رسیدن به یک جمع‌بندی نهایی از موضوع می‌باشد. از سوی دیگر، مشکلاتی مانند وجود ناسازگاری یا تضاد در بین داده‌ها و اطلاعات، بهنگام و تکمیل شدن این داده‌ها و اطلاعات در دوره‌های زمانی کوتاه، عدم اطمینان از مناسب و مرتبط بودن داده‌ها و اطلاعات و غیره نیز وجود دارد (ننکوا و مکون، ۲۰۱۲).

خلاصه‌سازی خودکار اسناد متنی برخط، راه‌حلی است که در سال‌های اخیر برای حل این مشکلات ارائه شده و به صورت گسترده مورد توجه محققان قرار گرفته است. با استفاده از مکانیزم‌های خلاصه‌سازی خودکار اسناد متنی می‌توان خلاصه‌ای از اطلاعات متنی برخط تولید کرد که کاربر با مطالعه این خلاصه، نسبت به محتوای اطلاعاتی اسناد مربوطه در مدت زمان کمی آگاهی پیدا کرده و بر اساس آن نسبت به استفاده یا عدم استفاده از آن اسناد به صورت مناسب تصمیم‌گیری کند.

یک سیستم خلاصه‌سازی خودکار اسناد متنی، سیستمی است که یک یا چند سند متنی را به عنوان ورودی دریافت کرده با انجام مجموعه‌ای از پردازش‌ها بر روی این ورودی وبدون دخالت کاربر انسانی، یک خلاصه منسجم و روان با طول مشخص و از پیش‌تعریف شده تولید می‌کند (ننکوا و مکون، ۲۰۱۲). منظور از خلاصه در این تعریف، متنی است که از یک یا چند سند متنی تولید شده و حاوی اطلاعات مهم موجود در اسناد

اصلی بوده و طول آن از نصف طول سند یا اسناد اصلی کوتاه‌تر است (رادو و هوی و مکون، ۲۰۰۲). در واقع، هدف در این سیستم‌ها تولید خلاصه‌ای است که با کمترین افزونگی، بیشترین پوشش از محتوای اطلاعاتی سند یا اسناد ورودی را داشته باشد. از سوی دیگر، خلاصه تولید شده به وسیله سیستم بایستی در قیاس با خلاصه‌های تولید شده توسط افراد خبره نیز حد قابل قبولی از کارایی، انطباق و خوانایی را داشته باشد به گونه‌ای که کاربران انسانی بتوانند با مطالعه آن، بر مهم‌ترین مطالب موجود در سند یا اسناد ورودی احاطه یابند (رانکل و کنروی و دنگ و ننکوا، ۲۰۱۳).

مقاله حاضر بر سیستم‌های خلاصه‌سازی چندسندی استخراجی مبتنی بر پرس‌وجو با رویکرد خوشه‌بندی مبتنی بر تفسیر متنی و با استفاده از استلزام متنی، متمرکز می‌باشد. در این راستا، مسئله خلاصه‌سازی به صورت یک مسئله بهینه‌سازی تعریف شده و از دو روش حریصانه و برنامه‌ریزی پویا برای حل آن استفاده شده است. ادامه مقاله به صورت زیر سازمان‌دهی شده است. بخش دوم به بررسی پیشینه پژوهش پرداخته، در بخش سوم سیستم پیشنهادی مطرح شده، در بخش چهارم نتایج حاصل از پیاده‌سازی و اجرای سیستم نمایش داده شده و مورد تحلیل قرار گرفته‌اند. در نهایت در بخش پنجم به جمع‌بندی و بیان راهکارهایی برای پژوهش‌های آتی پرداخته شده است.

## پیشینه پژوهش

اولین گام در سیستم‌های خلاصه‌سازی متن استخراج ویژگی‌های جمله‌های ورودی است. در گام بعد، امتیاز هر جمله بر اساس ترکیبی از امتیاز ویژگی‌های آن محاسبه شده و در نهایت، براساس امتیاز جمله‌ها، بهترین نمونه‌ها برای ساخت خلاصه نهایی انتخاب و استخراج می‌شوند (ننکوا و مکون، ۲۰۱۱). مهمترین ویژگی‌های مورد استفاده در چنین روش‌هایی عبارتند از تعداد تکرار کلمات (ننکوا و مکون ۲۰۱۲)، احتمال کلمات (لین و چن، ۲۰۱۰)، موقعیت کلمه (اوراسان و پکار و هسلر، ۲۰۰۴)، بکارگیری مرکز ثقل مجموعه اسناد (فیلاتوا و هتزیواسیلوگو، ۲۰۰۴) و مدل‌های بیزی (لی و لی، ۲۰۱۳). البته در برخی از این سیستم‌ها، از ترکیبی از این ویژگی‌ها با استفاده از تکنیک‌های مدل پنهان

مارکف (کونروی و اولیری، ۲۰۰۱)، شبکه‌های عصبی (ماریو و سعیدی و رودریگز و برانکو و سیلوا، ۲۰۱۷)، الگوریتم ژنتیک (لیتواک و لاست و فریدمن، ۲۰۱۰)، میدان تصادفی شرطی (شن و سان و لی و یانگ و چن، ۲۰۰۷) و الگوریتم‌های یادگیری ماشین با ناظر و بدون ناظر (هونگ و ننکوا، ۲۰۱۴) و (اسکلاتر و سوگارد، ۲۰۱۵) نیز استفاده شده است.

یکی از مشکلات روش‌های فوق این است که الگوریتم رده‌بندی تنها مشخص می‌کند آیا جمله‌ای برای ساختن خلاصه ارزشمند می‌باشد یا خیر. حال اگر تعداد جمله‌ها ارزشمند زیاد باشند، الگوریتم‌های فوق کمکی در انتخاب جمله‌های ارزشمند نمی‌کنند. برای حل این مشکل عمدتاً از الگوریتم‌های رتبه‌بندی مبتنی بر یادگیری استفاده شده است. در این روش‌ها، که نمونه‌ی آن در (کای و لی، ۲۰۱۳) وجود دارد، از یک الگوریتم تکمیلی برای رتبه‌بندی جمله‌های مناسب و در نهایت انتخاب جمله‌های دارای رتبه بهتر استفاده شده است. روش دیگر رتبه‌بندی جمله‌ها استفاده از روشهای رگرسیون می‌باشد. در این روش‌ها که از رگرسیون بردار پشتیبانی استفاده شده است، کیفیت رده‌بندی و رتبه‌بندی افزایش یافته است (اویانگ و لی و لی و لو، ۲۰۱۱).

با این حال، توجه به ارتباط معنایی بین جمله‌ها در اغلب روش‌های فوق نادیده گرفته شده است. در واقع، بسیاری از جمله‌هایی که در الگوریتم‌های رتبه‌بندی فوق‌الذکر در قالب جمله‌های ارزشمند رتبه‌بندی می‌شوند، از لحاظ معنایی به هم نزدیک بوده و وجود یکی از آنها در خلاصه نهایی کافی می‌باشد. به این ترتیب شناسایی ارتباط معنایی بین جمله‌ها از طریق روش‌هایی مانند استلزام متنی و استفاده از آن برای حذف جمله‌های دارای معنای نزدیک به هم، می‌تواند در کاهش فضای مصرفی خلاصه و امکان به کارگیری جمله‌های بیشتر در آن به منظور افزایش میزان پوشش خلاصه، بسیار موثر باشد.

## استلزام متنی

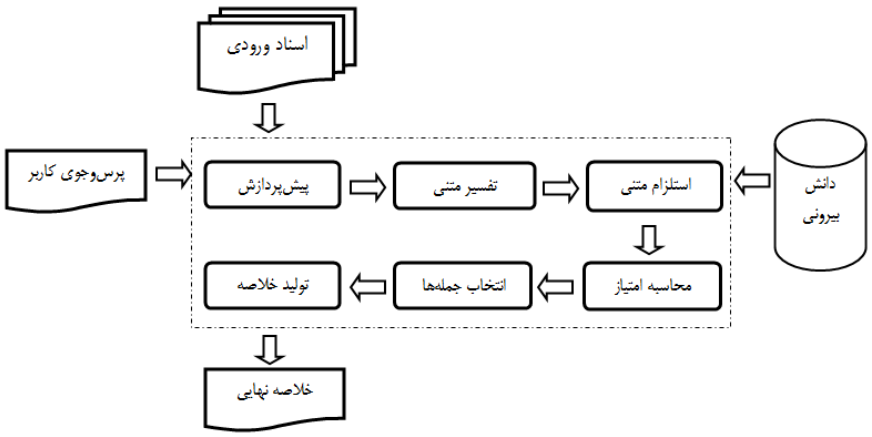
یکی از روش‌هایی که در دهه اخیر برای شناسایی ارتباط بین جمله‌ها در سیستم‌های خلاصه‌سازی متن مورد استفاده قرار گرفته است، استلزام متنی می‌باشد. استلزام متنی یک

رابطه یک طرفه بین یک متن (Text) و یک فرضیه متنی (Hypothesis) است که نشان می‌دهد آیا می‌توان فرضیه را از متن استنتاج کرد یا خیر. در واقع، هدف استلزام متنی آن است که با در نظر گرفتن یک متن (T) و با استفاده از دانش مرتبط با موضوع، مشخص کند آیا فرضیه از آن متن استلزام می‌گردد یا خیر. باید توجه داشت که نبایستی تنها بر دانش جنبی برای استلزام تکیه کرد و در این راستا، متن (T) نقش اصلی را بایستی ایفا کند. به عنوان مثال جدول شماره ۱ نمونه‌هایی از استلزام و عدم استلزام فرضیه از طریق متن را نمایش می‌دهد.

در (تاتار و تاماینو موریتا و میهایس و لوپسا، ۲۰۰۸) و (کوپتا و کاتوریا و ساین و ساچدوا و بهاتی، ۲۰۱۲) از استلزام متنی برای شناسایی جمله‌های با درجه ارتباط بالا استفاده شده است. همچنین، در (کوپتا و کور و میرکین و ساین و گویال، ۲۰۱۴) نیز از استلزام متنی و مسئله پوشش رأس کمینه وزن‌دار برای خلاصه‌سازی متن استفاده شده است. با این حال، روش‌های فوق همگی بر خلاصه‌سازی تک‌سندی متمرکز بوده و به دلیل بررسی رابطه استلزام متنی بین تمامی جمله‌ها، از لحاظ زمانی پیچیدگی مطلوبی ندارند.

**جدول ۱. نمونه‌هایی از استلزام متنی**

| Text | Hypothesis | Result |
|---|---|---|
| The two suspects belong to the 30th Street gang, which became embroiled in one of the most notorious recent crimes in Mexico: a shootout at the Guadalajara airport in May, 1993, that killed Cardinal Juan Jesus Posadas Ocampo and six others. | Cardinal Juan Jesus Posadas Ocampo died in 1993. | True |
| Regan attended a ceremony in Washington to commemorate the landings in Normandy. | Washington is located in Normandy. | False |
| Google files for its long awaited IPO. | Google goes public. | True |
| The SPD got just 21.5% of the vote in the European Parliament elections, while the conservative opposition parties polled 44.5%. | The SPD is defeated by the opposition parties. | True |
| According to NC Articles of Organization, the members of LLC company are H. Nelson Beavers, III, H. Chester Beavers and Jennie Beavers Stewart. | Jennie Beavers Stewart is a share-holder of Carolina Analytical Laboratory. | False |




**شکل ۱. چارچوب کلی روش پیشنهادی**

## روش پیشنهادی

شکل ۱ چارچوب کلی روش پیشنهادی در این مقاله را نمایش می‌دهد. هر یک از مؤلفه‌های این روش در ادامه مورد بررسی قرار خواهند گرفت.

## پیش‌پردازش

این مؤلفه وظیفه خواندن اسناد ورودی در قالب متن و پرس‌وجوی کاربر و انجام عملیات تقسیم‌بندی جمله‌ها، تجزیه لغوی، نرمال‌سازی، ریشه‌یابی، برچسب‌زنی اجزای کلام و حذف کلمات توقف را بر عهده دارد.

## تفسیر متنی

هدف از این مؤلفه، خوشه‌بندی جمله‌ها بر اساس امتیاز تفسیر و تأویل متنی بین آنها می‌باشد. به عبارت دیگر، در این مؤلفه همه جمله‌های که دارای امتیاز تفسیر و تأویل متنی بیش از یک آستانه مشخص شده هستند در یک خوشه قرار می‌گیرند. به این منظور به ازای هر جمله‌ی $S_1$ و $S_2$، امتیاز تفسیر متنی آن‌ها بر اساس رابطه ۱ محاسبه شده و جمله‌هایی که

امتیاز آن‌ها از مقدار آستانه از قبل تعریف شده (بر اساس مقدار به دست آمده در فاز آموزش) بیشتر باشد، در یک خوشه قرار می‌گیرند.

$$P(w_1, w_2) = \frac{1}{4}(PPDB_{2.0Score}(w_1, w_2) + Sim_{Distrib}(w_1, w_2) + Sim_{PPDB.Cos}(w_1, w_2) + Sim_{PPDB.JS}(w_1, w_2)) \quad \text{رابطه ۱)}$$

به طوری که $w_1$ به کلمه‌های جمله‌ی $S_1$ و $w_2$ به کلمه‌های جمله‌ی $S_2$ اشاره داشته و مقدار $PPDB_{2.0Score}(w_1,w_2)$ با استفاده از رابطه‌ی تعریف شده در (پاولیک و راستوگی و گانیت کویچ و ون دورم و کالیسون بورچ، ۲۰۱۵) به دست می‌آید. به علاوه، بر اساس بردارهای ویژگی $v_i^p$ تعریف شده در (کوکوس و کالیسون بورچ، ۲۰۱۶) مقادیر *SimPPDB.Cos* و *SimPPDB.JS* به کمک روابط ۲ و ۳ به دست می‌آیند.

$$\mathrm{Sim}_{\mathrm{PPDB.Cos}}(i, j) = \cos(v_i^p, v_j^p) \quad \text{رابطه ۲)}$$

$$\mathrm{Sim}_{\mathrm{PPDB.JS}}(i, j) = 1 - \mathrm{JS}(v_i^p, v_j^p) \quad \text{رابطه ۳)}$$

به طوری که، مقدار JS بر اساس مقدار واگرایی جنسن–شانون به دست می‌آید و در محاسبه آن فرض بر نرمال بودن توزیع احتمال کلمه $i$ بر اساس $v_i^p$ می‌باشد.

همچنین مقدار *SimDistrib* بر اساس رابطه ۴ به دست می‌آید.

$$\mathrm{Sim}_{\mathrm{Distrib}}(i, j) = \cos(v_i^w, v_j^w) \quad \text{رابطه ۴)}$$

به طوری که $v_i^w$ عبارت است از بردار ۳۰۰ بعدی *WORD2VEC* حاصل از نمایش کلمه i که به وسیله مجموعه‌داده‌های گوگل نیوز آموزش داده شده (میکولو و ساتسکور و چن و کورادو و دیین، ۲۰۱۳). هم‌چنین تابع *cos* بر اساس رابطه شباهت کسینوسی تعریف شده در رابطه ۵ محاسبه می‌شود.

$$\mathrm{Cos}(q, s) = \frac{\sum_i (tf * idf)_{i,q} * (tf * idf)_{i,s}}{\sqrt{\sum_{i \in m} ((tf * idf)_{i,q})^2} * \sqrt{\sum_{i \in n} ((tf * idf)_{i,s})^2}} \quad \text{رابطه ۵)}$$

به طوری که مقادیر tf و idf به ترتیب بر اساس روابط ۶ و ۷ به دست می‌آیند.

$$\mathrm{tf}(a, s) = 0.5 + 0.5 \times \frac{f_{a,s}}{\max_{\{a' \in s\}} f_{a',s}} \quad \text{رابطه ۶)}$$

که $f_{a,s}$ فراوانی عبارت a در جمله s را نشان می‌دهد.

رابطه ۷)
$$\mathrm{idf}(a, S) = \log \frac{N}{|\{s \in S: a \in s\}|}$$

به طوری که N برابر با تعداد کل اسناد ورودی می‌باشد.

به این ترتیب، با اتمام فرآیند خوشه‌بندی جمله‌ها در خوشه‌هایی قرار می‌گیرند که از لحاظ معنایی با یکدیگر ارتباط نزدیکی دارند. سپس امتیاز هر خوشه در رابطه با پرس‌وجوی مطرح شده بر اساس رابطه ۱ (بر مبنای جمله‌های درون هر خوشه و جمله‌های تشکیل دهنده پرس‌وجو) محاسبه شده و جمله سرآیند درصد مشخصی از خوشه‌ها که بیشترین امتیاز را داشته باشند، برای ادامه فرایندهای روش پیشهادی انتخاب می‌شوند. لازم به ذکر است که سرآیند هر خوشه، جمله‌ای است که بیشترین ارتباط با پرس‌وجو (بر اساس رابطه ۱) را داشته باشد.

## استلزام متنی

هدف از این مؤلفه، بررسی رابطه استلزام متنی بین همه جمله‌ها حاصل از مؤلفه قبل، به منظور امتیازدهی به جمله‌ها و رتبه‌بندی آنها می‌باشد. به عبارت دیگر، در این مؤلفه تمامی جمله‌ها به صورت دوتایی‌های متن و فرضیه تعریف شده و ضمن بررسی رابطه استلزام متنی بین آنها، ضریب اطمینان این رابطه نیز به دست آمده و بر اساس این ضرایب، امتیاز هر جمله به دست می‌آید. به این منظور، ابتدا هر یک از جمله‌ها به عنوان متن لحاظ شده و ضریب اطمینان حاصل از بررسی استلزام متنی به ازای فرضیه بودن سایر جمله‌ها در ارتباط با آن بررسی می‌شود. به ازای هر فرضیه، ضریب اطمینان مربوطه در سطر مربوط به متن و ستون مربوط به آن فرضیه، در ماتریس $TEM$ ذخیره می‌گردد. لازم به ذکر است که ضریب اطمینان یک رابطه استلزام متنی عبارت است از عددی بین صفر تا یک که بیشتر بودن آن به معنای قطعیت بیشتر رابطه است. به این ترتیب، مقادیر ماتریس $TEM$ به ازای هر دو جمله $s_1$ و $s_2$ بر اساس رابطه ۸ تعریف می‌شوند.

رابطه (۸)
$$TEM(s_1, s_2) = TE\_Conf(Text = s_1, Hyp = s_2)$$

به طوری که *TE_Conf* ضریب اطمینان حاصل از بررسی رابطه استلزام متنی با فرض متن بودن جمله $s_1$ و فرضیه بودن جمله $s_2$ می‌باشد.

## محاسبه امتیاز

هدف از این مؤلفه، محاسبه امتیاز نهایی هر جمله به منظور انتخاب جمله‌ها بهتر برای ساخت خلاصه می‌باشد. به این منظور، امتیاز هر جمله مانند $s$ بر اساس امتیاز حاصل از بررسی روابط استلزام متنی در گام قبل یعنی ماتریس TEM، و با استفاده از رابطه ۹ به دست می‌آید.

$$Final\_Score(s) = \begin{cases} +\infty & if \ \sum_x TEM[x][s] = 0 \\ \frac{\sum_x TEM[s][x]}{\sum_x TEM[x][s]} & Otherwise \end{cases}$$

رابطه ۹)

## انتخاب جمله‌ها

بعد از محاسبه امتیاز نهایی جمله‌ها، بایستی بهترین جمله‌ها برای ساخت خلاصه نهایی انتخاب شوند. به این منظور، با توجه به محدود بودن طول خلاصه مطلوب، بایستی جمله‌های انتخاب شوند که ضمن رعایت این محدودیت، الزامات پوشش حداکثری محتوای اطلاعاتی اسناد ورودی و مطابقت خلاصه نهایی با نیازهای کاربر (پرس‌وجوی مطرح شده) را نیز برآورده کنند. در چنین شرایطی، مسئله خلاصه‌سازی اسناد متنی را می‌توان در قالب یک مسئله بهینه‌سازی فرموله‌سازی کرده و از طریق حل آن، خلاصه نهایی را تولید کرد.

به این ترتیب، می‌توان مسئله خلاصه‌سازی را به صورت پیدا کردن زیرمجموعه‌ای مانند A از مجموعه جمله‌ها موجود فرموله‌سازی کرد که رابطه‌ی ۱۰ را برآورده کند.

$$\max \left( \sum_{s \in A} Final_{Score(s)} \right) \quad subject\ to \quad \sum_{s \in A} l_s \leq L$$

رابطه ۱۰)

به طوری که در آن، مقدار $Final\_Score$ بر اســاس رابطه ۹ به دســت آمده و $l_s$ و L به ترتیب طول جمله s و طول خلاصه مطلوب می‌باشند.

در نهایت، جمله‌ها انتخاب شــده در این مؤلفه برای تولید خلاصــه‌ی نهایی در اختیار آخرین مؤلفه سیستم قرار می‌گیرند.

## تولید خلاصه

هدف از این مؤلفه تولید خلاصـه نهایی بر اسـاس جمله‌ها انتخاب شـده توسـط مؤلفه قبل می‌باشـد. به این منظور ابتدا جمله‌ها بر اسـاس ترتیب حضـور آن‌ها در اسـناد ورودی و همچنین ترتیب زمانی مرتب می‌شــوند. به عبارت دیگر، جمله‌های که در اســناد ورودی زودتر ظاهر شده‌اند، در خلاصه نهایی تولید شده نیز جلوتر از سایر جمله‌ها قرار می‌گیرند. همچنین در صورتی که بتوان رابطه زمانی (بر اساس پارامترهایی مانند قیدهای زمان، تاریخ یا تأخر و تقدم معنایی) بین جمله‌ها پیدا کرد، از چنین رابطه‌ای نیز برای مرتب‌سازی دقیق‌تر جمله‌ها استفاده می‌شود. در انتها، از روش‌های حل مرجع ضمیر برای مشخص کردن مرجع ضمایر در خلاصه نهایی استفاده می‌شود. لازم به ذکر است به دلیل تغییر جمله‌ها در مراحل قبل و به منظور جلوگیری از خطا، این فرآیند بر روی جمله‌های اصــلی مرتبط با جمله‌های انتخاب شــده اعمال شــده و ســپس نتیجه آن با جمله‌های موجود در خلاصــه نهایی منطبق می‌شود. به این ترتیب و با اتمام فرآیندهای این مؤلفه، خلاصه‌نهایی سیستم پیشنهادی تولید شده و به عنوان خروجی سیستم ارائه می‌شود. جدول شماره ۲ بخش از خلاصه تولید شده توسط سیستم پیشنهادی را نشان می‌دهد.

**جدول (۲): بخش از خلاصه‌ی تولید شده برای مجموعه‌ی شماره ۱۸ از DUC 2007**

| **Query: How has Starbucks Coffee attempted to expand and diversify through joint ventures, acquisitions, or subsidiaries?** |
|---|
| Generated Summary: Seattle-based Starbucks hopes that with Tazo it can attract new customers. Steve Smith, who founded Tazo in 1994, would not disclose what Starbucks paid for Tazo. By expanding Starbucks' online sales beyond coffee, Schultz said the company is taking a crack at potential revenues estimated at $100 billion. Starbucks alone reported $1.1 billion in retail sales for fiscal 1998. Schultz said that Starbucks would set up a portal site by the end of this year. Starbucks coffeehouses are already all over the place. Starting this spring, the service also will deliver Starbucks by the pound. …. |

## یافته‌های پژوهش

### جزییات پیاده‌سازی

برای پیاده‌سازی سیستم پیشنهادی، از جعبه‌ابزار زبان طبیعی پایتون و در مؤلفه استلزام متنی از ابزار EOP 1.2.3 استفاده شده است. این ابزار با ۴۰۰ زوج متن و فرضیه از RTE-5 و ۱۰۰ زوج تولید شده به صورت دستی از مجموعه داده‌های مورد استفاده، آموزش داده شده است. همچنین در مؤلفه‌ی انتخاب جمله‌ها، مسئله‌ی مورد نظر در قالب مسئله‌ی کوله‌پشتی صفر و یک فرموله‌سازی شده و با استفاده از دو رویکرد حریصانه و برنامه‌ریزی پویا حل شده است که به ترتیب تضمین کننده سرعت و دقت در پاسخ می‌باشند.

### مجموعه داده‌ها

برای ارزیابی دقیق نتایج حاصل از اجرای سیستم و ایجاد قابلیت مقایسه نتایج با سایر سیستم‌ها، از مجموعه داده‌های مجموعه داده DUC 2007 (فرایند اصلی) به عنوان مجموعه داده‌های آموزش و آزمایش استفاده شده است. این مجموعه‌داده‌ها دارای ۴۵ مجموعه داده می‌باشد که هر مجموعه شامل یک موضوع، ۲۵ سند متنی مرتبط با موضوع و یک پرس‌وجو می‌باشد. از این مجموعه داده به صورت تصادفی از ۳۰ مجموعه داده برای آموزش و از ۱۵ مجموعه باقیمانده برای آزمایش سیستم استفاده شده است. طول خلاصه مورد نظر در این مجموعه‌ها ۲۵۰ کلمه است.

### معیارهای ارزیابی

برای ارزیابی نتایج نیز از روش ارزیابی خودکار بر مبنای سیستم ROUGE استفاده شده است. این سیستم به صورت خودکار از طریق مقایسه خلاصه‌های تولید شده توسط یک سیستم خلاصه‌سازی اسناد متنی با خلاصه‌های استاندارد طلایی (خلاصه‌ای که توسط یک انسان خبره برای همان اسناد ورودی نوشته شده است) بر مبنای معیارهای متفاوت، کیفیت و کارایی آن‌ها را در قالب سه پارامتر دقت، بازخوانی و F-Measure گزارش می‌دهد. در بین معیارهای متفاوتی که از طریق سیستم ROUGE قابل اندازه‌گیری هستند، سه معیار ROUGE-1، ROUGE-2 و ROUGE SU-4 به دلیل ارتباط معنایی نزدیک‌تر

آن‌ها را با سیستم مبتنی بر قضاوت انسانی و استفاده بیشتر در گزارش نتایج پژوهش‌های مرتبط، به منظور نمایش و گزارش نتایج حاصل از اجرای سیستم پیشنهادی بر روی مجموعه داده‌های مورد نظر انتخاب شده‌اند.

### سیستم‌های مبنا

به منظور مقایسه دقیق نتایج حاصل از اجرای سیستم با سیستم‌های مشابه، از بهترین سیستم‌های شرکت کننده در مجموعه داده، میانگین نتایج شرکت‌کنندگان در مجموعه داده و بهترین سیستم پیشنهاد شده در صورت وجود اطلاعات مناسب، به عنوان سیستم‌های مبنا استفاده شده است.

### نتایج ارزیابی

جدول شماره ۳ نتایج ارزیابی خودکار سیستم پیشنهادی در مقایسه با بهترین سیستم‌های مشارکت‌کننده در مجموعه داده‌های DUC 2007 را نشان می‌دهد. همچنین جدول شماره ۴ نتایج ارزیابی سیستم پیشنهادی در مقایسه با سایر سیستم‌های مبنا را نمایش می‌دهد. در هر دو جدول بهترین مقدار برای مقایسه بهتر به صورت تو پر نشان داده شده‌اند. به علاوه، در این دو جدول نتایج سیستم پیشنهادی بر اساس رویکردهای حریصانه و برنامه‌ریزی پویا گزارش شده‌اند.

**جدول ۳. مقایسه نتایج سیستم پیشنهادی با بهترین سیستم‌های مشارکت کننده در DUC 2007**

| سیستم | ROUGE-1 F-Measure | ROUGE-2 F-Measure | ROUGE-SU4 F-Measure |
|---|---|---|---|
| سیستم ۲۵ | ۰/٤٤ | ۰/۱۲ | ۰/۱۷ |
| سیستم ۱۵ | ۰/٤٤ | ۰/۱۲ | ۰/۱۷ |
| میانگین | ۰/٤۰ | ۰/۱۰ | ۰/۱۵ |
| حریصانه | ۰/٤٦ | **۰/۱۳** | ۰/۱۷ |
| برنامه‌ریزی پویا | **۰/٤۷** | **۰/۱۳** | **۰/۱۹** |

## تحلیل نتایج

همان‌گونه که نتایج جدول‌های ۲ و ۳ نشان می‌دهند، سیستم پیشنهادی نسبت به سایر سیستم‌های مبنا (سیستم‌های مشارکت‌کننده در مجموعه داده و سایر سیستم‌های مبنا) از نتایج بهتری برخوردار است. این امر به طور ویژه در رابطه با رویکرد برنامه‌ریزی پویا صدق

کرده که مطابق انتظار نیز بوده است. این موارد نشان می‌دهند سیستم پیشنهادی از کیفیت مناسبی برای خلاصه‌سازی اسناد متنی برخوردار است.

همچنین، در سنجه ROUGE-2 رویکرد حریصانه به طرز قابل توجهی نسبت به برنامه‌ریزی پویا نتیجه مناسب‌تری را تولید کرده است. دلیل این امر آن است که مطابق نظر (لین و هوی، ۲۰۰۳) سنجه ROUGE-2 نزدیکترین سنجه کارایی به تفکرات انسانی است و از آنجا که رویکرد حریصانه نیز در بین رویکردهای حل مسئله بیشترین پیوند را با نحوه حل مسئله توسط انسان دارد، این نتیجه چندان دور از انتظار و نامعقول نیست و حتی می‌توان آن را دلیل برای اثبات ادعای (لین و هوی، ۲۰۰۳) نیز دانست.

**جدول ٤. مقایسه نتایج سیستم پیشنهادی با سایر سیستم‌های مبنا**

| سیستم | ROUGE-1 F-Measure | ROUGE-2 F-Measure | ROUGE-SU4 F-Measure |
|---|---|---|---|
| Conroy(2007) | ۰/٤۳ | ۰/۱۰ | ۰/۱٦ |
| Toutanova(2007) | – | ۰/۰۹ | ۰/۱۳ |
| Haghighi (2009) | ۰/٤۳ | ۰/۱۰ | ۰/۱٥ |
| Mason(2011) | ۰/٤۳ | ۰/۱۲ | ۰/۱۷ |
| He (2012) | ۰/٤۰ | ۰/۰۷ | – |
| Wang (2013) | – | **۰/۱۳** | ۰/۱۸ |
| Cai (2013) | ۰/٤۲ | ۰/۱۲ | – |
| Li (2015) | ۰/٤۰ | ۰/۰۹ | ۰/۱٥ |
| Canhasi (2016) | ۰/٤۲ | ۰/۰۹ | ۰/۱٦ |
| Cao (2016) | ۰/٤٤ | ۰/۱۲ | – |
| Feigenblat (2017) | ۰/٤۳ | ۰/۱۱ | ۰/۱۷ |
| Chali (2017) | – | **۰/۱۳** | ۰/۱۳ |
| Mani (2017) | ۰/٤۲ | ۰/۱۱ | ۰/۱۷ |
| حریصانه | ۰/٤٦ | **۰/۱۳** | ۰/۱۷ |
| برنامه‌ریزی پویا | **۰/٤۷** | **۰/۱۳** | **۰/۱۹** |

## نتیجه‌گیری و پیشنهادها

این مقاله روشی جدید برای خلاصه‌سازی استخراجی چندسندی متون ارائه کرده است. در این مقاله، مسئله خلاصه‌سازی به صورت مسئله بهینه‌سازی کوله‌پشتی فرموله‌سازی

شـده و با اسـتفاده از رویکردهای تفسـیر و اسـتلزام متنی ضـمن خوشـه‌بندی جمله‌ها، امتیاز آن‌ها محاسبه شده، سپس با استفاده از روش‌های حریصانه و برنامه‌ریزی پویا نسبت به حل مسـئله اقدام شـده اسـت. این روش به طرز چشـمگیری نسـبت به تمامی روش‌های مشـابه کارایی بیشـتری داشـته اسـت، به طوری که در مقایسـه با سـیسـتم‌های مشـارکت کننده در مجموعه داده‌های DUC 2007، به میزان ۰/۰۳ و در مقایسه با سایر سیستم‌های مبنا به میزان ۰/۰۲ عملکرد خلاصه‌سازی را بهبود داده است.

برای تحقیقات آتی علاقمند هستیم که رویکرد خوشه‌بندی اولیه جمله‌ها را با افزودن ویژگی‌هایی مانند موقیت جمله تقویت کرده و همچنین از روش‌های فشـرده‌سـازی جمله برای کـاهش طول جملـه‌هـا اسـتفـاده کنیم. بـه علاوه، بـه کـارگیری این روش روی خلاصه‌سازی بهنگام و همچنین سیستم‌های پرسش و پاسخ از دیگر اهداف آینده این مقاله می‌باشند.

## منابع


- Nenkova, A., & McKeown, K. (2012). A survey of text summarization techniques. In Mining text data (pp. 43-76). Springer, Boston, MA.
- Radev, D. R., Hovy, E., & McKeown, K. (2002). Introduction to the special issue on summarization. Computational linguistics, 28(4), 399-408.
- Rankel, P. A., Conroy, J. M., Dang, H. T., & Nenkova, A. (2013). A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (Vol. 2, pp. 131-136).
- Nenkova, A., & McKeown, K. (2011). Automatic summarization. Foundations and Trends® in Information Retrieval, 5(2–3), 103-233.
- Lin, S. H., & Chen, B. (2010, July). A risk minimization framework for extractive speech summarization. In Proceedings of the 48th annual meeting of the Association for Computational Linguistics (pp. 79-87). Association for Computational Linguistics.
- Orasan, C., Pekar, V., & Hasler, L. (2004, May). A Comparison of Summarisation Methods Based on Term Specificity Estimation. In LREC.
- Filatova, E., & Hatzivassiloglou, V. (2004, August). A formal model for information selection in multi-sentence text extraction. In Proceedings of the 20th international conference on Computational Linguistics (p. 397). Association for Computational Linguistics.
- Li, J., & Li, S. (2013). A novel feature-based bayesian model for query focused multi-document summarization. Transactions of the Association for Computational Linguistics, 1, 89-98.
- Conroy, J. M., & O'leary, D. P. (2001, September). Text summarization via hidden markov models. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 406-407). ACM.
- Maraev, V., Saedi, C., Rodrigues, J., Branco, A., & Silva, J. (2017, September). Character-level convolutional neural network for paraphrase detection and other experiments. In Conference on Artificial Intelligence and Natural Language (pp. 293-304). Springer, Cham.
- Litvak, M., Last, M., & Friedman, M. (2010, July). A new approach to improving multilingual summarization using a genetic algorithm. In Proceedings of the 48th annual meeting of the association for computational linguistics (pp. 927-936). Association for Computational Linguistics.
- Shen, D., Sun, J. T., Li, H., Yang, Q., & Chen, Z. (2007, January). Document summarization using conditional random fields. In IJCAI (Vol. 7, pp. 2862-2867).
- Hong, K., & Nenkova, A. (2014). Improving the estimation of word importance for news multi-document summarization. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (pp. 712-721).
- Natalie Schluter and Anders Sogaard, "Unsupervised extractive summarization via coverage maximization with syntactic and semantic concepts", Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 840–844, Beijing, China, July 26-31, 2015.
- Cai, X., & Li, W. (2013). Ranking through clustering: An integrated approach to multi-document summarization. IEEE Transactions on Audio, Speech, and Language Processing, 21(7), 1424-1433.
- Ouyang, Y., Li, W., Li, S., & Lu, Q. (2011). Applying regression models to query-focused multi-document summarization. Information Processing & Management, 47(2), 227-237.
- Tatar, D., Tamaianu-Morita, E., Mihis, A., & Lupsa, D. (2008). Summarization by logic segmentation and text entailment. Advances in Natural Language Processing and Applications, 15, 26.
- Gupta, A., Kathuria, M., Singh, A., Sachdeva, A., & Bhati, S. (2012, December). Analog textual entailment and spectral clustering (atesc) based summarization. In International Conference on Big Data Analytics (pp. 101-110). Springer, Berlin, Heidelberg.
- Gupta, A., Kaur, M., Mirkin, S., Singh, A., & Goyal, A. (2014). Text summarization through entailment-based minimum vertex cover. In Proceedings of the Third Joint Conference on Lexical and Computational Semantics (* SEM 2014) (pp. 75-80).
- Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B., & Callison-Burch, C. (2015). PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification.

In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (Vol. 2, pp. 425-430).
- Cocos, A., & Callison-Burch, C. (2016). Clustering paraphrases by word sense. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1463-1472).
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119)
- Wang, L., Raghavan, H., Castelli, V., Florian, R., & Cardie, C. (2016). A sentence compression based framework to query-focused multi-document summarization. arXiv preprint arXiv:1606.07548.
- Haghighi, A., & Vanderwende, L. (2009, May). Exploring content models for multi-document summarization. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 362-370). Association for Computational Linguistics.
- Cai, X., & Li, W. (2013). Ranking through clustering: An integrated approach to multi-document summarization. IEEE Transactions on Audio, Speech, and Language Processing, 21(7), 1424-1433.
- Toutanova, K., Brockett, C., Gamon, M., Jagarlamudi, J., Suzuki, H., & Vanderwende, L. (2007). The pythy summarization system: Microsoft research at duc 2007. In Proc. of DUC (Vol. 2007).
- Chali, Y., Tanvee, M., & Nayeem, M. T. (2017). Towards Abstractive Multi-Document Summarization Using Submodular Function-Based Framework, Sentence Compression and Merging. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (Vol. 2, pp. 418-424).
- Canhasi, E. and I. Kononenko, Weighted hierarchical archetypal analysis for multi-document summarization. Computer Speech & Language, 2016. 37: pp. 24-46.
- Cao, Z., Li, W., Li, S., Wei, F., & Li, Y. (2016). Attsum: Joint learning of focusing and summarization with neural attention. arXiv preprint arXiv:1604.00125.
- Conroy, J. M., Schlesinger, J. D., & O'Leary, D. P. (2007). Classy 2007 at duc 2007. In Proceedings of the Document Understanding Conference 2007.
- Feigenblat, G., Roitman, H., Boni, O., & Konopnicki, D. (2017, August). Unsupervised query-focused multi-document summarization using the cross entropy method. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 961-964). ACM.
- He, Z., Chen, C., Bu, J., Wang, C., Zhang, L., Cai, D., & He, X. (2012, July). Document summarization based on data reconstruction. In Twenty-Sixth AAAI Conference on Artificial Intelligence.
- Li, P., Bing, L., Lam, W., Li, H., & Liao, Y. (2015, June). Reader-aware multi-document summarization via sparse coding. In Twenty-Fourth International Joint Conference on Artificial Intelligence
- Mani, K., Verma, I., Meisheri, H., & Dey, L. (2018, December). Multi-document summarization using distributed bag-of-words model. In 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI) (pp. 672-675). IEEE.
- Mason, R., & Charniak, E. (2011, June). Extractive multi-document summaries should explicitly not contain document-specific content. In Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages (pp. 49-54). Association for Computational Linguistics.
- Lin, C. Y., & Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics.

**استناد به این مقاله:** شناسه دیجیتال (DOI): 10.22091/jemsc.2018.1270

ناصراسدی، علی. (۱۳۹۷). «خلاصه‌سازی چندسندی استخراجی مبتنی بر پرس‌وجوی متن با استفاده از تفسیر و استلزام متنی». *مدیریت مهندسی و رایانش نرم*، ۶ (۲)، ۱۹۸-۱۸۳.