An Automatic Persian Text Summarization System Based on Linguistic Features and Regression

Mahmood Soltani
M.Sc. in Computer Engineering; Instructor;
Department of Computer Engineering;
Quchan University of Technology m.soltani@qiet.ac.ir

Jalal Nasiri
PhD in Computer Engineering; Assistant Professor;
Iranian Research Institute for Information Science and Technology (IranDoc);
Corresponding Author j.nasiri@irandoc.ac.ir

Ehsan Asgarian
PhD in Computer Engineering; Engineering Department of Ferdowsi University of Mashhad ehsan.asgarian@mail.um.ac.ir

Iranian Journal of Information Processing and Management

Received: 14. Dec. 2016 **Accepted: 14, Jan. 2017**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 33 | No. 4 | pp. 1809-1828
Summer 2018

Abstract: Considering the vast amount of existing written information and the shortage of time, optimal summarization of books, articles, news reports, etc. on the Web is a major concern of researchers. In this paper, we propose a new approach for Persian single-document summarization based on several linguistic features of text. In our approach after extracting the linguistic features for each sentence, the weight of features is learned by a linear regression method. We select one sentence with maximum score at each step of algorithm. The score of each sentence is calculated based on two factors: first, sum of the weighted features and second, the amount of its similarity to the sentences that are selected for final summary previously. We use an automatic evaluation tool to compare our approach with other existing approaches. The result indicates that our method improves the performance of summarization.

Keywords: Single-Document Summarization, Linguistic Feature, Linear Regression

ارائهٔ سیستم خلاصه‌ساز متون فارسی بر مبنای ویژگی‌های زبان‌شناختی و رگرسیون

محمود سلطانی
کارشناسی ارشد؛ مهندسی کامپیوتر؛ مربی؛
دانشگاه صنعتی قوچان m.soltani@qiet.ac.ir

جلال‌الدین نصیری
دکتری؛ مهندسی کامپیوتر؛ استادیار؛
پژوهشگاه علوم و فناوری اطلاعات (ایرانداک)؛
پدیدآور رابط j.nasiri@Irandoc.ac.ir

احسان عسکریان
دکتری؛ مهندسی کامپیوتر؛ پژوهشگر؛ دانشگاه فردوسی مشهد ehsan.asgarian@mail.um.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۵/۱۱/۲۰ **پذیرش:** ۱۳۹۶/۰۶/۲۸ مقاله برای اصلاح به مدت ۴۱ روز نزد پدیدآوران بوده است.

فصلنامه | علمی پژوهشی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۳ | شمارهٔ ٤ | ص‌ص ۱۸۰۹-۱۸۲۸
تابستان ۱۳۹۷

چکیده: گسترش روزافزون داده‌های متنی فارسی در فضای اینترنت و پیچیدگی جست‌وجو در میان انبوه این اسناد، خلاصه‌سازی خودکار متون فارسی را به یکی از زمینه‌های تحقیقاتی مورد توجه تبدیل کرده است. در این مقاله روشی کارا برای خلاصه‌سازی خودکار متون فارسی ارائه شده است. روش پیشنهادی که به‌صورت انتخابی و تک‌سندی است، خلاصه‌سازی را بر اساس رتبه‌بندی جملات و انتخاب مهم‌ترین آن‌ها انجام می‌دهد. اهمیت هر جمله از متن با ترکیب خطی مقادیر هفت ویژگی زبان‌شناختی مستخرج از سند برای هر جمله به‌دست می‌آید. وزن بهینهٔ هر ویژگی در این ترکیب از روش رگرسیون خطی و با استفاده از پیکرهٔ آموزشی پاسخ محاسبه شده است. پس از محاسبهٔ اهمیت جملات متن، در هر مرحله از الگوریتم، یک جمله با اهمیت بیشتر تا رسیدن به نرخ فشرده‌سازی مورد نظر انتخاب می‌شود. این جمله علاوه بر این‌که دارای بیشترین اهمیت است، کمترین میزان شباهت با جملات انتخاب‌شده در مراحل قبلی را نیز دارد. نتایج به‌دست‌آمده از مقایسهٔ الگوریتم پیشنهادی با دو سیستم خلاصه‌ساز «ایجاز» و «فارسی‌سام» با استفاده از «پیکرهٔ پاسخ» نشان می‌دهد که در بیشتر معیارهای ارزیابی پیشرفت قابل توجهی حاصل شده است.

کلیدواژه‌ها: خلاصه‌سازی تک‌سندی، ویژگی‌های زبان‌شناختی متن، رگراسیون خطی

۱. مقدمه

رشد سریع وب و منابع دادهٔ موجود در آن باعث شده است که کاربران آن با حجم بسیار بالایی از اطلاعات در زمینه‌های مختلف مواجه باشند. بر همین اساس، پیدا کردن اطلاعات مفید در این حجم انبوه داده برای انسان کاری بسیار سخت، وقت‌گیر و در برخی موارد غیرممکن است. یکی از روش‌های مقابله با این مشکل خلاصه‌سازی خودکار متون است. خلاصه‌سازی متون می‌تواند به ساده شدن کار و همچنین، کاهش زمان جست‌وجو کمک کند. ایدهٔ اصلی در خلاصه‌سازی خودکار پیدا کردن یک نسخهٔ مختصرتر از سند اصلی توسط ماشین است، به‌طوری که مفهوم اصلی و ویژگی‌های مهم سند اولیه حفظ شود. خلاصه‌سازی خودکار یکی از زمینه‌های تحقیقاتی مهم در پردازش زبان‌های طبیعی است که می‌تواند در بسیاری موارد از جمله متن‌کاوی، سیستم‌های پرسش و پاسخ خودکار، جمع‌آوری خلاصهٔ اخبار، بازیابی اطلاعات و ... کاربرد داشته باشد.

روش‌های خلاصه‌سازی خودکار متون را می‌توان به دو دستهٔ استخراجی[1] و چکیده‌ای[2] تقسیم کرد. در خلاصه‌سازی استخراجی، ساختار جملات متن اصلی تغییر نمی‌کند. در این روش‌ها جملات متن اصلی بر اساس ویژگی‌های مشخص وزن‌دهی شده و جملات با وزن بیشتر به‌عنوان جملات مهم‌تر در خلاصهٔ نهایی انتخاب می‌شوند. در مقایسه با این نوع خلاصه‌سازی، در خلاصه‌سازی چکیده‌ای معمولاً ساختار کلی جملات در سند اصلی تغییر می‌کند. خلاصه‌سازی در این دسته شبیه به خلاصه‌های انسانی است. با توجه به پیچیدگی این دسته از خلاصه‌سازها، اغلب روش‌های موجود از نوع استخراجی است. از جنبه‌ای دیگر، خلاصه‌سازی به دو دستهٔ تک‌سندی[3] و چندسندی[4] تقسیم‌بندی می‌شود، به‌طوری که در خلاصه‌سازی تک‌سندی، یک سند به‌عنوان سند اولیه خلاصه می‌شود و در خلاصه‌سازی چندسندی، خلاصه‌ای از چندین سند مرتبط از لحاظ موضوعی تهیه می‌شود و هدف، تهیهٔ یک نسخهٔ خلاصه از ترکیب این اسناد است. پیچیدگی‌ها در روش‌های چندسندی بیشتر از روش‌های تک‌سندی است، چرا که ممکن است چندین سند با ارتباط موضوعی اما با دیدگاه‌های متفاوت و حتی متناقض داشته باشیم.

روش‌های اولیهٔ خلاصه‌سازی متون از ویژگی‌های سادهٔ آماری مانند تعداد تکرار

1. extractive 2. abstractive 3. single-document summarization
4. multi-document summarization

کلمات، طول، و موقعیت قرار گرفتن جمله در متن استفاده می‌کردند. این در حالی است که در سال‌های اخیر استفاده از ابزارهای جدید در زمینهٔ پردازش متن، به‌کارگیری روش‌های یادگیری ماشین و همچنین، تحلیل‌های معنایی توانسته است تا حدود زیادی کیفیت سیستم‌های خلاصه‌ساز را افزایش دهد. در اغلب سیستم‌های خلاصه‌سازِ خودکار از ویژگی‌های ظاهری متن از جمله بسامد کلمات، موقعیت جملات، عبارات خاص و تشابه با عنوان متن برای ورودی الگوریتم‌های یادگیری ماشین مانند روش «بیز»[1]، درخت تصمیم، مدل مخفی «مارکوف»[2]، شبکه‌های عصبی و ... استفاده شده است.

با توجه به تعداد زیاد روش‌های پیشنهادی به‌منظور خلاصه‌سازی خودکار متون باید روشی برای ارزیابی و مقایسهٔ این روش‌ها وجود داشته باشد. در سال‌های گذشته تلاش‌های بسیاری برای این منظور صورت گرفته است. برای نمونه، می‌توان به انجمن DUC[3] اشاره کرد که به‌منظور ارزشیابی سیستم‌های خلاصه‌ساز، هر ساله با در اختیار قرار دادن پیکره‌ها و دادگان مناسب به برگزاری همایش و رقابت می‌پردازد (Harman and Over 2002). در این میان معیارهای مختلفی برای ارزشیابی و مقایسهٔ الگوریتم‌های خلاصه‌ساز ارائه شده است. در این زمینه ROUGE مجموعه‌ای از معیارها و نرم‌افزارهایی است که برای ارزشیابی سیستم‌های خلاصه‌ساز خودکار و ترجمهٔ ماشینی استفاده می‌شود (Lin 2004). متأسفانه برای زبان فارسی پیکره‌های محدودی برای ارزشیابی و یادگیری سیستم‌های خلاصه‌ساز وجود دارد. «پیکرهٔ پاسخ» یکی از این نمونه‌هاست که دارای ۱۰۰ متن از اخبار با موضوعات مختلف به همراه پنج خلاصهٔ انسانی برای هر کدام است (Moghaddas et al. 2013). در این مقاله از «پیکرهٔ پاسخ» به‌منظور یادگیری سیستم و ارزشیابی نهایی آن استفاده شده است.

ساختار این مقاله در ادامه به‌صورت زیر است: بخش بعدی به معرفی مسئله و جنبه‌های مختلف آن می‌پردازد. در بخش سوم، به بررسی کارهای مرتبط پرداخته شده است. در بخش چهارم، روش پیشنهادی و مراحل الگوریتم خلاصه‌ساز معرفی شده است. در بخش پنجم، علاوه بر معرفی معیارهای ارزشیابی و جزئیات آزمایش‌های انجام‌شده بر روی سیستم، خلاصه‌ساز پیشنهادی ارائه و با دیگر روش‌ها مقایسه شده و در نهایت، در بخش ششم، نتیجه‌گیری انجام می‌گیرد.

1. Bayes 2. Markov 3. Document Understanding Conference

۲. معرفی مسئله

مسئله‌ای که در این مقاله به آن پرداخته شده، خلاصه‌سازی خودکار متون خبری است. خلاصه‌سازی خودکار به فرایند تجمیع و فشرده‌سازی یک یا چند منبع اطلاعاتی به‌منظور کوتاه‌کردن متن و یا پاسخگویی به درخواست کاربر بدون دخالت انسان گفته می‌شود. همان‌طور که در این تعریف آمده، خلاصه‌سازی از جنبه‌های مختلف قابل بررسی و دسته‌بندی است. «ننکووا و مک‌کوون» به بررسی این زوایا پرداخته است که در ادامه به برخی از آن‌ها اشاره می‌شود.

یکی از موضوعات قابل بررسی، هدف از تهیهٔ نسخهٔ خلاصه است. بر این اساس، تهیهٔ خلاصه می‌تواند برای پاسخگویی به پرس‌وجویی از سوی کاربر باشد که خلاصهٔ نهایی باید تا حد امکان به این درخواست نزدیک‌تر باشد. علاوه بر این خلاصه‌سازی می‌تواند در راستای به‌روزرسانی پایگاه داده و افزودن جمع‌بندی متون دیگر در قالب یک متن خلاصه باشد. جنبهٔ قابل بررسی دیگر تعداد منابعی است که باید بر اساس آن‌ها متنِ خلاصه تهیه شود. خلاصه‌سازی یا بر روی یک متن ورودی صورت می‌پذیرد، که به آن خلاصه‌سازی تک‌سندی گفته می‌شود و یا باید یک نسخهٔ خلاصه از چندین متن موجود تهیه شود، که به آن خلاصه‌سازی چندسندی می‌گویند. روش تهیهٔ خلاصه یکی دیگر از جنبه‌های تفاوت سیستم‌های خلاصه‌ساز است. خلاصه‌سازی می‌تواند به‌صورت استخراجی باشد که در آن متن خلاصه، زیربخشی از متن اصلی است. به‌عبارت دیگر، گزیده‌ای از متن اصلی بدون تغییر ساختاری در متن خلاصه ظاهر می‌شود. نوع دیگری که به مراتب چالش‌برانگیزتر از نوع قبلی است، خلاصه‌سازی چکیده‌ای است. در این نوع خلاصه‌سازی، جملاتی با ساختار جدید توسط ماشین و برگرفته از متن اصلی ساخته می‌شود که علاوه بر کوتاه‌تر بودن نسبت به متن اصلی، مفهوم کلی آن را بیان می‌کند. همان‌طور که دیده می‌شود، دراین نوع خلاصه‌سازی جنبه‌های دستوری و نگارشی زبان مورد نظر باید در نظر گرفته شود.

با توجه به جنبه‌های مختلف خلاصه‌سازی که در بالا بیان گردید، در این مقاله تمرکز به روی خلاصه‌سازی تک‌سندی به‌روش استخراجی است.

۳. کارهای مرتبط

اولین تلاش‌ها برای خلاصه‌سازی خودکار متون، به کار Luhn (1958) برمی‌گردد

که در آن، اطلاعات آماری از بسامد کلمات و توزیع آن‌ها در متن مبنای خلاصه‌سازی قرار گرفته است. از آن به بعد روش‌های بسیاری بر اساس اطلاعات آماری، زبان‌شناسی، معنایی و با رویکردهای یادگیری ماشین ارائه شده که بسیاری از آن‌ها به‌صورت تک‌سندی و استخراجی هستند. به‌عنوان مثال، «هوی و لین» روشی بر اساس امتیازدهی به جملات و انتخاب جملات مهم ارائه کرده‌اند (Hovy and Lin 1998). در ادامه، با توجه به فراوانی تعداد روش‌های خلاصه‌سازی متون فقط به بررسی برخی از روش‌ها در زبان فارسی پرداخته شده است.

در سال‌های اخیر تلاش‌های زیادی برای خلاصه‌سازی خودکار متون فارسی انجام شده است (Shamsfard, Akhavan, and Jourabchi 2009; Kiyoumarsi, Eslami, and Tajoddin 2010; Zamanifar, Minaei-Bidgoli, and Sharifi 2008; Hassel, Nada, and Mazdak 2004; Pour Masoomi et al. 2014 و ... . «شمس‌فرد، اخوان و جورابچی» در این سیستم‌ها به‌منظور امتیازدهی به جملات از چندین ویژگی متنی استفاده کرده‌اند. در هر مرحله از الگوریتم جمله‌ای که به جملات انتخاب‌شده در مراحل قبل بیشترین شباهت را دارد، حذف می‌شود و از جملات باقی‌مانده، جملهٔ با اهمیت بیشتر انتخاب می‌شود (Shamsfard, Akhavan, and Jourabchi 2009). در روش «کیومرثی، اسلامی و تاج‌الدین» با استفاده از «وردنت» و منطق فازی مناسب‌ترین جملات به‌عنوان جملات خلاصهٔ نهایی انتخاب شده‌اند (Kiyoumarsi, Eslami, and Tajoddin 2010). در روش ارائه‌شده در کار «زمانی‌فر، مینایی-بیدگلی و شریفی» با استفاده از هم‌رخداد کلمات برای پیداکردن کلمات مهم و همچنین، با استفاده از مترادف کلمات برای جلوگیری از انتخاب جملات مشابه، به خلاصه‌سازی متون پرداخته شده (Zamanifar, Minaei-Bidgoli, and Sharifi 2008) که نتایج آن در مقایسه با کار Hassel, Nada, and Mazdak (2004) بهتر گزارش شده است.

الگوریتمی که در کار «هاسل، ندا و مزدک» پیشنهاد شده، تنها بر اساس اطلاعات آماری مستخرج از متن است که در واقع، نسخهٔ فارسی از الگوریتم SwuSum است (Hassel, Nada, and Mazdak 2004). این الگوریتم برای خلاصه‌سازی متون زبان سوئدی مورد استفاده قرار می‌گیرد. «ایجاز»، یکی دیگر از سیستم‌های خلاصه‌ساز خودکار برای زبان فارسی است که در آن از ۹ ویژگی متنی به‌منظور ارزش‌گذاری جملات استفاده شده است. با استفاده از روش کمترین، مربعات وزن هر ویژگی محاسبه شده و در نهایت، جملات با بیشترین درجهٔ اهمیت انتخاب می‌شوند. این سیستم به‌صورت تک‌سندی و

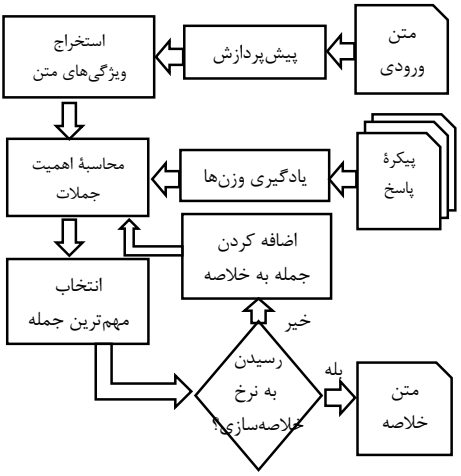
چندسندی می‌تواند خلاصهٔ مناسب را تولید نماید (Pour Masoomi et al. 2014). «زمانی‌فر و کاشفی» از ویژگی‌های آماری و مفهومی متن استفاده کرده و پس از تبدیل ساختار متن ورودی به درخت فراکتال، خلاصه‌ای از آن را تولید می‌کنند (Zamanifar and Kashefi 2011). یکی دیگر از روش‌ها، روشی است که در آن از هستان‌شناسی استفاده می‌شود. در الگوریتم پیشنهادی از مجموعهٔ واژگان «فارس‌نت»[1] به‌منظور پیدا کردن رابطهٔ معنایی بین بخش‌های مختلف متن استفاده شده است. این روابط در یک گراف مبتنی بر متن پیاده‌سازی شده و در نهایت، از روی آن جملات مهم برای خلاصهٔ نهایی انتخاب می‌شوند (Ramezani and Feizi-Derakhshi 2015).

همان‌طور که دیده می‌شود، شیوهٔ کار در اغلب روش‌های بررسی‌شده، خلاصه‌سازی استخراجی و تک‌سندی است. تعدادی از آن‌ها بر پایهٔ اطلاعات آماری و ساختاری متن و تعدادی بر اساس اطلاعات مفهومی و هستان‌شناسی‌ها بنا شده‌اند. علاوه بر این، فرایند انتخاب جملات مهم در آن‌ها متفاوت است. روش‌های مبتنی بر نظریهٔ گراف‌ها، منطق فازی، دسته‌بندی و خوشه‌بندی نمونه‌هایی از روش‌های مورد استفاده است.

٤. روش پیشنهادی

در این مقاله روشی به‌منظور خلاصه‌سازی خودکار متون فارسی ارائه شده که بر مبنای خلاصه‌سازی تک‌سندی و استخراجی است. به‌عبارت دیگر، جملات مهم متن ورودی را شناسایی و با توجه به نرخ فشرده‌سازی مهم‌ترین آن‌ها را انتخاب می‌کند. شکل ۱، مراحل کلی الگوریتم را نشان داده است.

1. FarsNet



شکل ۱. مراحل انجام الگوریتم پیشنهادی

الگوریتم پیشنهادی در این مقاله دارای چهار مرحلهٔ اصلی است: پیش‌پردازش متن ورودی، استخراج مقادیر ویژگی‌های هفت‌گانه، محاسبهٔ اهمیت هر جمله بر اساس ویژگی‌ها، و انتخاب جملات مهم برای خلاصهٔ نهایی. علاوه بر این، قبل از شروع الگوریتم یک پردازش اولیه به‌منظور یادگیری وزن هر ویژگی بر روی «پیکرهٔ پاسخ» انجام می‌شود. در ادامه، به بررسی بیشتر هر یک از این مراحل پرداخته شده است:

٤-١. پیش‌پردازش متن

به‌علت وجود برخی چالش‌های نگارشی و دستوری در زبان فارسی (Shamsfard 2011)، انجام برخی پیش‌پردازش‌ها بر روی متن اولیه برای رسیدن به دقت بیشترِ خلاصه‌سازی لازم و ضروری است. در این فاز، چهار اقدام اصلی وجود دارد: جداسازی جملات، جداسازی کلمات، حذف کلمات توقف[1]، و ریشه‌یابی کلمات. در مرحلهٔ جداسازی، با استفاده از علائم جداکننده (.، ؟، :، و ...) و یک‌سری قواعد خاص، جملات متن ورودی استخراج می‌شود. در ادامه، با توجه به ویژگی‌های دستوری زبان فارسی، کلمات هر جمله جدا شده و پس از آن کلمات پرتکرار و بی‌اهمیت که کلمات توقف نامیده می‌شوند (مانند: از، که، است، و ...) از لیست کلمات حذف می‌شوند. ریشه‌یابی کلمات آخرین

1. stop list

مرحله از پیش‌پردازش است که در آن ریشهٔ کلمات با کلمهٔ اصلی جایگزین می‌شود. این کار به‌منظور افزایش دقت در تحلیل‌های آماری و بسامد کلمات است.

٤-۲. استخراج ویژگی‌های متنی

در خلاصه‌سازی استخراجی، مبنا انتخاب مهم‌ترین جملات متن برای خلاصهٔ نهایی است. به‌منظور تشخیص اهمیت هر جمله باید معیارها و ویژگی‌هایی وجود داشته باشد تا بر اساس آن‌ها بتوان میزان اهمیت هر جمله را محاسبه کرد. در این مقاله از هفت ویژگی پُرکاربرد در مقالات مشابه استفاده شده است که لیست آن در جدول ۱، آورده شده است. این فاکتورها اطلاعات آماری و زبان‌شناختی نهفته در متن را جمع آوری می‌کنند. در این مقاله مقدار هر ویژگی پس از محاسبه نرمال‌سازی شده و مقداری بین صفر و یک به هر کدام تخصیص داده می‌شود. در ادامه، به معرفی این ویژگی‌ها پرداخته شده است.

جدول ۱. ویژگی‌های استخراج‌شده برای هر جمله

نماد	توضیحات
F1	تشابه جمله با عنوان متن اصلی
F2	طول جمله بر اساس تعداد کلمات آن
F3	شباهت متن با زمینهٔ متن اصلی
F4	موقعیت جمله در متن
F5	وجود عبارات خاص در جمله
F6	وجود اسامی خاص
F7	وجود اعداد

٤-۲-۱. تشابه با عنوان

استفاده از عنوان یکی از ابتدایی‌ترین شاخص‌ها برای انتخاب جملات متن خلاصه است. میزان شباهت هر جمله با عنوان متن بر اساس روش‌های سنجش میزان شباهت جملات محاسبه می‌شود. معمولاً جملاتی در متن که با عنوان آن تشابه بیشتری دارند، جملات مهم‌تری هستند. این مفهوم یکی از معیارهایی است که برای محاسبهٔ اهمیت یک جمله در این مقاله استفاده شده است. روش محاسبهٔ این معیار در فرمول (۱) آورده شده است.

$$\mathrm{sim}(T,S_i)=\frac{1}{\mathrm{hamming\ distance}(T,S_i)}$$

$$F1_i=\frac{\mathrm{sim}(T,S_i)}{\mathrm{Max}_k(\mathrm{sim}(T,S_k))} \quad (۱)$$

که در آن $\mathrm{sim}(T,S_i)$ شباهت بین عنوان و جملهٔ S_i است.

٤-۲-۲. طول جملات

جملات خیلی کوتاه و یا خیلی طولانی معمولاً جملات مفیدی برای متن خلاصه نیستند. به‌طور کلی، در جملات کوتاه اطلاعات مفیدی دیده نمی‌شود و جملات طولانی با توجه به ذاتِ خلاصه‌سازی نمی‌توانند انتخاب مناسبی باشند. می‌توان با تعریف یک حد آستانه برای طول جمله، جملاتی را که در این محدوده قرار نگیرند، حذف کرد.

البته، در روش پیشنهادی هیچ جمله‌ای به این علت حذف نمی‌شود و تنها این معیار مانند معیارهای دیگر در محاسبهٔ ارزش جملات به‌کار گرفته می‌شود. به‌عبارت دیگر، تصمیم‌گیری برای انتخاب و یا عدم انتخاب یک جمله به تمام ویژگی‌های آن بستگی دارد و ممکن است ویژگی‌های دیگر، اهمیت جمله‌ای با طول زیاد و یا طول کم را بالا برده و الگوریتم آن جمله را برای متن خلاصه انتخاب کند. نحوهٔ محاسبهٔ این معیار در فرمول (۲) آورده شده است.

$$F2_i=\frac{|Si|}{\mathrm{Max}_k(|S_k|)} \quad (۲)$$

که در آن $|s|$ طول جمله s و یا تعداد کلمات آن است.

٤-۲-۳. شباهت با زمینهٔ متن

جملاتی که به زمینهٔ اصلی متن شبیه‌ترند، می‌توانند گزینهٔ مناسب‌تری برای انتخاب در خلاصهٔ نهایی باشند. زمینهٔ اصلی متن را می‌توان از روی کلمات پرتکرار به‌دست آورد. کلمات پرتکرار در متن در واقع، مشخص‌کنندهٔ زمینهٔ متن هستند. بنابراین، جملاتی که شامل تعداد بیشتری از این کلمات پرتکرار باشند، به زمینهٔ متن نزدیک‌تر هستند. فرمول (۳) به‌منظور محاسبهٔ شباهت جملات با زمینهٔ متن اصلی استفاده شده است.

$$F3_i = \frac{\sum_{j=1}^{l} \mathrm{tf}(w_j^i)}{\mathrm{Max}_k(\sum_{j=1}^{l} w_j^k)} \quad (۳)$$

که در آن $tf(w_j^i)$ بسامد کلمهٔ j در جملهٔ i و l طول جمله است.

٤-۲-٤. موقعیت جمله

جایگاه جملات در متن تا حدود بسیار زیادی می‌تواند روی اهمیت آن‌ها تأثیرگذار باشد. در زبان انگلیسی تحقیقات زیادی ثابت کرده است که جملات ابتدایی در متون خبری از اهمیت بالایی برخوردار هستند. در زبان فارسی نیز این موضوع تا حدود بسیار زیادی برقرار است. هرچند که در برخی موارد به‌علت عدم رعایت اصول نگارشی این موضوع تحت تأثیر قرار می‌گیرد، با این حال، در این مقاله از فرمول (۴) به‌منظور محاسبهٔ تأثیر جایگاه جملات استفاده شده است.

$$F4_i = \frac{1}{\mathrm{pos}(S_i)} \quad (۴)$$

که در آن $\mathrm{pos}(S_i)$ موقعیت جملهٔ iام در متن است.

٤-۲-٥. عبارات خاص

وجود برخی عبارات خاص در جملات اهمیت ویژه‌ای به جمله می‌دهد. در مقابل، عباراتی هستند که وجود آن‌ها در جمله از اهمیت آن کم می‌کنند. برای محاسبهٔ نقش این عبارات در اهمیت جمله، از پنج نوع عبارت خاص استفاده شده است: عبارات مهم، عبارات غیرمهم، القاب، ضمایر و علائم نقل قول. به‌عنوان مثال، جمله‌ای که شامل عبارت «در نتیجه» باشد به احتمال زیاد حاوی اطلاعات مفید است و در مقابل، جمله‌ای که در آن از عبارت «به‌عنوان مثال» استفاده شده است، نمی‌تواند گزینهٔ خوبی برای متن خلاصه باشد. جدول ۲، شامل برخی از این عبارات خاص است. در فرمول (۵) نحوهٔ محاسبهٔ تاثیر عبارات خاص در اهمیت جمله آورده شده است.

$$sp(S_i) = \mathrm{im}(S_i) + \mathrm{h}(S_i) + \mathrm{qu}(S_i) - \mathrm{uim}(S_i) - \mathrm{pr}(S_i)$$
$$F5_i = \frac{\mathrm{sp}(S_i) - \mathrm{Min}_k(\mathrm{sp}(S_k))}{\mathrm{Max}_k(\mathrm{sp}(S_k)) - \mathrm{Min}_k(\mathrm{sp}(S_k))} \quad (۵)$$

که در آن $\mathrm{im}(S_i)$ تعداد عبارات مهم، $\mathrm{h}(S_i)$ تعداد القاب، $\mathrm{uim}(S_i)$ تعداد عبارات غیرمهم، $\mathrm{pr}(S_i)$ تعداد ضمایر و $\mathrm{qu}(S_i)$ تعداد علائم نقل قول جملهٔ i است.

جدول ۲. مثال‌هایی از عبارات خاص در جملات

نوع عبارات	مثال
عبارت مهم	در نتیجه، بنابراین، در مجموع
عبارت غیرمهم	مثلا، به‌عنوان مثال، مانند
القاب	آقای، خانم، بانو، سید، سیده، استاد
ضمیر	آن‌ها، او
علائم نقل قول	" "

٤-٢-٦. اسامی خاص

معمولاً وجود اسامی خاص در جمله می‌تواند به اهمیت آن بیفزاید. بر این اساس، جمله‌ای که شامل تعداد بیشتری اسم خاص باشد، می‌تواند اهمیت بیشتری داشته باشد. این ویژگی در متون خبری که به بیان یک خبر می‌پردازد، بیشتر دیده می‌شود. این در حالی است که وجود اسامی خاص در متونی که ارجاعات زیادی به اسامی دارد، می‌تواند نقش منفی در خلاصه‌سازی داشته باشد. فرمول (۶) برای محاسبهٔ مقدار این ویژگی استفاده می‌شود.

$$F6_i = \frac{\mathrm{NE}(S_i)}{\mathrm{Max}_k\left(\mathrm{NE}(S_k)\right)} \qquad (۶)$$

که در آن $\mathrm{NE}(S_i)$ تعداد اسامی خاص در جملهٔ S است.

٤-٢-٧. اطلاعات عددی

معمولاً جملاتی دارای اطلاعات عددی مانند تاریخ، درصدها و ... از اهمیت بیشتری برخوردار هستند. فرمول (۷) نرمال‌شدهٔ تعداد اطلاعات عددی در جمله است.

$$F7_i = \frac{\mathrm{ND}(S_i)}{\mathrm{Max}_k(\mathrm{ND}(S_k))} \qquad (۷)$$

که در آن $ND(S_i)$ تعداد اطلاعات عددی در جمله است.

٤-٣. محاسبهٔ میزان اهمیت هر جمله

در این مرحله، برای هر جمله هفت مقدار بر اساس پارامترهای توضیح داده‌شده در بخش قبل به‌دست آمده است که باید بر اساس آن‌ها میزان اهمیت هر جمله را محاسبه

کـرد. نقـش هـر پارامتـر در تعییـن اهمیـت جملـه یکسـان نیسـت. به‌عبـارت دیگـر، ترکیـب پارامترهـا بـا وزن یکسـان به‌خوبـی گویـای اهمیـت جمـلات نیسـت. بـرای محاسـبهٔ وزن هـر پارامتـر در تعییـن اهمیـت جملـه می‌تـوان از روش‌هـای یادگیـری ماشـین اسـتفاده نمـود.

یکـی از ایـن روش‌هـا رگرسـیون خطـی اسـت. مـا از نرم‌افـزار WEKA بـرای یادگیـری وزن‌هـا بـر اسـاس ایـن روش اسـتفاده کرده‌ایـم. بـرای ایـن منظـور، از ۲۰۰۰ جملـه «پیکـرهٔ پاسـخ» اسـتفاده شـده اسـت. بـا توجـه بـه این‌کـه در «پیکـرهٔ پاسـخ» هـر متـن خلاصـه دارای پنـج خلاصـهٔ انسـانی اسـت، بـه هـر جملـه بسـته بـه این‌کـه در چنـد خلاصـهٔ انسـانی ظاهـر شـده اسـت، یـک عـدد بیـن صفـر تـا پنـج انتسـاب داده می‌شـود (C). بنابرایـن، بـه جملـه‌ای کـه در هیـچ خلاصـهٔ انسـانی وجـود نـدارد، عـدد صفـر و بـه جملـه‌ای کـه در تمـام خلاصه‌هـای انسـانی یـک متـن دیـده شـده اسـت، عـدد پنـج تخصیـص داده می‌شـود.

جـدول ۳، بخشـی از دادگان یادگیـری را نشـان می‌دهـد کـه بـرای هـر جملـه مقـدار هفـت ویژگـی و عـدد انتسـاب داده‌شـده بـه آن (C) آورده شـده اسـت.

جدول ۳. بخشی از دادگان آموزشی

شمارهٔ جمله	F1	F2	F3	F4	F5	F6	F7	تعداد خلاصهٔ انسانی (C)
۱	۰/۷۵	۰/۴۹	۰/۶۳	۰/۱۲	۰	۰	۰/۲۳	۰
۲	۰/۸	۰/۵	۰/۶۴	۰/۱۱	۰/۵	۰	۰/۲۹	۲
۳	۰/۶۶	۰/۴۳	۱	۰/۱	۰/۳۳	۰	۰/۲۹	۱
۴	۰/۸۲	۰/۵۱	۰/۸۴	۰/۰۹	۰/۵	۰	۰/۲۹	۳
۵	۰/۸۴	۰/۵۱	۰/۵۹	۰/۰۸	۰/۵	۰	۰/۴۱	۳

٤-٤. انتخاب جملات متن خلاصه

هـدف اصلـی در خلاصه‌سـازی اسـتخراجی انتخـاب جمـلات مهـم اسـت. یکـی از روش‌هـای به‌دسـت آوردن جمـلات مهـم، انتسـاب یـک مقـدار عـددی بـه هـر جملـه به‌عنـوان میـزان اهمیـت و انتخـاب جمـلات بـا بیشـترین ایـن مقـدار اسـت. مـا در روش پیشـنهادی بـرای به‌دسـت آوردن میـزان اهمیـت هـر جملـه از ترکیـب هفـت ویژگـی استخراج‌شـده از متـن اسـتفاده کرده‌ایـم. ایـن هفـت ویژگـی و نحـوهٔ اسـتخراج آن‌هـا در قسـمت‌های قبـل توضیـح داده شـد. در ایـن بخـش، نحـوهٔ محاسـبهٔ میـزان اهمیـت جمـلات از روی ایـن ویژگـی شـرح داده می‌شـود.

پس از پیش‌پردازش‌های لازم بر روی متن ورودی، مقادیر هفت ویژگی برای هر جمله استخراج می‌شود. پس از آن و بر اساس وزن هر ویژگی که با استفاده از روش رگرسیون خطی به‌دست آمده، با استفاده از فرمول (۸) می‌توان میزان اهمیت جملات را محاسبه کرد. در جدول ۴، این شیوهٔ امتیازدهی، روش اولیه نامیده شده است.

$$\text{Score}(S_i)=\frac{\sum_{k=0}^{n} F_k * W_k}{n} \quad (۸)$$

که در آن n تعداد ویژگی‌ها، F_k مقدار ویژگی k و W_k وزن ویژگی k است.

یکی از نقاط ضعف این شیوه برای محاسبهٔ میزان اهمیت جملات عدم توجه به میزان شباهت جملات انتخابی در خلاصهٔ نهایی است. با این شیوهٔ امتیاز دهی، در صورتی که دو جملهٔ شبیه به هم در متن اصلی امتیاز بالایی کسب کنند، هر دو جمله برای متن خلاصه انتخاب می‌شوند. این انتخاب باعث ایجاد افزونگی در متن خلاصه می‌شود. برای جلوگیری از این مشکل، روش امتیازدهی جدیدی بر اساس بیشترین شباهت حاشیه‌ای ارائه شده است.

(۹)

$$MMR(C'Q'R'S) = Argmax_{D_i \in R \setminus S}\left[\lambda * sim(D_i,Q) - (1-\lambda) * Max_{D_j \in S}\left(sim(D_i,D_j)\right)\right]$$

بیشترین شباهت حاشیه‌ای: در اغلب سیستم‌های بازیابی اطلاعات، لیستی از اسناد مرتبط با درخواست کاربر آماده می‌شود که بر اساس میزان تشابه سند با درخواست کاربر و همچنین، کمترین شباهت با اسناد قبلی مرتب می‌شود. ترکیب خطی این دو فاکتور شباهت حاشیه‌ای نامیده می‌شود. بنابراین، سندی با شباهت حاشیه‌ای بیشتر رتبهٔ بالاتری در لیست اسناد بازیابی‌شده دارد. فرمول (۹) نحوهٔ محاسبهٔ شباهت حاشیه‌ای را نشان می‌دهد که در آن c مجموعهٔ کل اسناد، Q درخواست کاربر، R لیست امتیاز دهی‌شده از اسناد بازیابی‌شده و S زیرمجموعهٔ R است که قبلاً برای کاربر آماده شده است.

در امتیازدهی جملات متن می‌توان از معیار شباهت حاشیه‌ای استفاده نمود. بنابراین، جمله‌ای بیشترین امتیاز را دارد که اولاً مقدار محاسبه‌شده برای آن از فرمول (۸) بیشترین بوده، و دوماً دارای کمترین شباهت با جملات انتخاب‌شدهٔ قبلی باشد. فرمول (۱۰) بیان‌کنندهٔ نحوهٔ محاسبهٔ امتیاز هر جمله بر اساس شباهت حاشیه‌ای است. در جدول ۴، این شیوهٔ امتیازدهی، روش ترکیبی نامیده شده است.

$$New\_Score(S_i \in Q) = \lambda * Score(S_i) - (1-\lambda) * Max_k(sim(S_i,R_k)) \quad (۱۰)$$

که در آن Q کل جملات متن اصلی، R زیرمجموعه‌ای از Q که قبلاً برای متن خلاصه انتخاب شده‌اند، k تعداد جملاتی که تا این مرحله انتخاب شده‌اند و $sim(S_i,R_k)$ میزان شباهت دو جمله است که در بخش‌های قبل توضیح داده شد.

در فرمول (۱۰) در صورتی که مقدار $\lambda = 1$ انتخاب شود، مقدار محاسبه‌شده با مقدار محاسبه‌شده از طریق فرمول (۸) یکسان است و اگر $\lambda = 0$ در نظر گرفته شود، این معیار فقط عدم تشابه را در نظر می‌گیرد، به‌طوری که جمله‌ای با کمترین تشابه با جملات قبلی بیشترین امتیاز را دارد. برای مقادیر دیگر λ بین بازهٔ صفر و یک، ترکیب خطی از دو مقدار در محاسبهٔ میزان اهمیت جمله در نظر گرفته می‌شود.

برای به‌دست آوردن مناسب‌ترین مقدار λ الگوریتم را بر مبنای مقادیر متفاوت λ و بر روی دادگان آزمایشی پاسخ اجرا کرده و بر اساس نتایج به‌دست آمده بهترین مقدار $\lambda = 0.6$ انتخاب گردید. جدول (۴) نتایج حاصل را نشان می‌دهد. همان‌طور که دیده می‌شود، مقدار $\lambda = 0.6$ در اقلب پارامترهای ارزیابی سبب ایجاد نتایج بهتری شده است. پارامترهای ارزیابی p در بخش بعدی مورد بررسی قرار گرفته‌اند.

در الگوریتم پیشنهادی در هر مرحله یکی از جملات متن اولیه که بیشترین امتیاز را دارد، انتخاب می‌شود. انتخاب جملات تا زمانی ادامه پیدا می‌کند که به نرخ فشرده‌سازی برسیم. به‌عبارت دیگر، تعداد کلمات متن خلاصه به درصد از پیش تعیین‌شده‌ای از تعداد کل کلمات متن اصلی برسد. فرمول (۱۱) نحوهٔ محاسبهٔ نرخ فشرده‌سازی را نشان می‌دهد.

جدول ٤. نتایج اجرای الگوریتم برای مقادیر مختلف λ

	λ						
P	۰/۸	۰/۷	۰/۶	۰/۵	۰/۴	۰/۳	۰/۲
P1	۰/۳۴۱	۰/۳۵۱	۰/۳۵۴	۰/۳۵۲	۰/۳۳۲	۰/۳۱۹	۰/۲۳۵
P2	۰/۳۵۱	۰/۳۶۱	۰/۳۶۴	۰/۳۶۱	۰/۳۴۰	۰/۳۲۸	۰/۳۳۴
P3	۰/۵۱۸	۰/۵۲۴	۰/۵۲۱	۰/۵۱۷	۰/۴۹۳	۰/۴۸۰	۰/۴۸۴
P4	۰/۲۶۰	۰/۲۶۷	۰/۲۶۹	۰/۲۷۴	۰/۲۵۹	۰/۲۵۱	۰/۲۵۷
P5	۰/۵۰۴	۰/۵۰۹	۰/۵۱۲	۰/۵۱۵	۰/۵	۰/۴۹۶	۰/۵۰۲
P6	۰/۳۸۴	۰/۳۹۳	۰/۳۹۶	۰/۳۹۴	۰/۳۷۴	۰/۳۶۴	۰/۳۷۰
P7	۰/۳۷۵	۰/۳۸۴	۰/۳۸۷	۰/۳۸۶	۰/۳۶۷	۰/۳۵۶	۰/۳۶۲

$$Compressionrate = \frac{lengthofsummary}{lengthoforiginal} \quad (۱۱)$$

در اولین مرحله از الگوریتم با توجه به این‌که در فرمول (۹) مجموعهٔ R شامل جمله‌ای نیست، مبنای محاسبهٔ امتیاز جملات فرمول (۸) است. در مراحل بعدی، ابتدا بر اساس فرمول (۹) امتیاز جملات باقی‌مانده مجدد محاسبه شده و جمله‌ای که امتیاز بیشتر را کسب کرده، به لیست جملات متن خلاصه افزوده می‌شود. در ادامه، الگوریتم با توجه به فرمول (۱۱) نرخ فشرده‌سازی تا این مرحله را محاسبه کرده و با مقدار از پیش تعیین‌شدهٔ قبلی مقایسه می‌کند. در صورتی که نرخ فشرده‌سازی از این مقدار کمتر باشد و یا به‌عبارت دیگر، هنوز خلاصهٔ نهایی کامل نشده باشد، الگوریتم ادامه پیدا کرده و پس از محاسبهٔ مجدد امتیاز جملات انتخاب‌نشده، مراحل تا رسیدن به نرخ فشرده‌سازی مناسب ادامه پیدا می‌کند.

۵. ارزیابی

برای ارزیابی سیستم پیشنهادی از یک سیستم ارزیاب خودکار برای زبان فارسی استفاده شده است و نتایج حاصل از ارزیابی با دو سیستم خلاصه‌ساز زبان (ایجاز و فارسی‌سام) مقایسه شده است. علاوه بر این، پیکرهٔ مورد استفاده برای ارزیابی «پیکرهٔ پاسخ» انتخاب شده است. پاسخ اولین پیکرهٔ متنی برای ارزیابی سیستم‌های خلاصه‌ساز فارسی است که در دو بخش تک‌سندی و چندسندی تهیه شده است. در بخش تک‌سندی، ۱۰۰ متن از اخبار در شش حوزهٔ مختلف (اقتصادی، اجتماعی، فرهنگی، ورزشی، سیاسی، و علمی) به همراه پنج متن خلاصه برای هر کدام وجود دارد.

در ارزیابی خلاصه‌سازی از چندین معیار استفاده می‌شود. در ادامهٔ به بررسی برخی از آن‌ها که در این مقاله استفاده شده، پرداخته شده است.

۵-۱. Nتایی‌های مشترک

این معیار تعداد N تایی‌های مشترک بین متن خلاصهٔ ماشینی و مجموعهٔ متون خلاصهٔ مرجع را محاسبه می‌کند که هم می‌تواند بر اساس جملهٔ (P1) و هم کل متن (P2) در نظر گرفته شود. زمانی که بر اساس جمله این پارامتر محاسبه می‌شود، کلمات دو جملهٔ متفاوت در یک Nتایی قرار نمی‌گیرند.

۵-۲. طولانی‌ترین زیررشتهٔ مشترک (LCS)

در این معیار ترتیب کلمات نیز مهم است و طولانی‌ترین کلمات متوالی مشترک بین دو متن مورد توجه قرار می‌گیرد. این معیار با عنوان P3 در جداول نمایش داده شده است.

۵-۳. تعداد جفت کلمات مشترک با فاصلهٔ آزاد

به هر جفت کلمه (با حفظ ترتیب) در جمله skip-bigram گفته می‌شود. این معیار با اندازه‌گیری تعداد skip-bigramهای مشترک بین خلاصهٔ ماشینی و خلاصهٔ مرجع محاسبه می‌شود.این معیار با عنوان P4 در جداول نمایش داده شده است.

۵-٤. تعداد کلمات مشترک

به‌عنوان ساده‌ترین معیار شباهت متن، می‌توان تعداد کلمات مشترک آن‌ها را به‌دست آورد. این معیار با عنوان P5 در جداول نمایش داده شده است.

۵-۵. معیار ویژهٔ Nتایی‌های مشترک

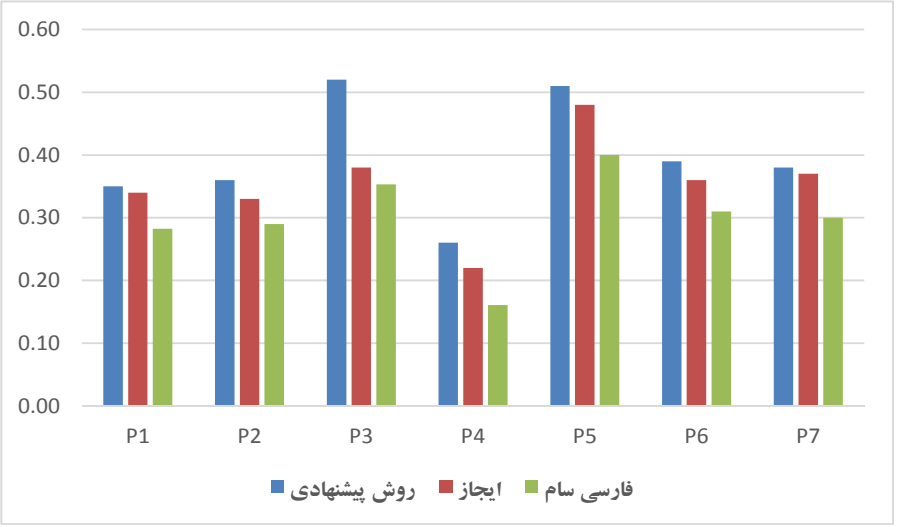
در این معیار علاوه بر محاسبهٔ تعداد Nتایی‌های مشترک تمام iتایی‌های مشترک که در آن‌ها i<N است نیز محاسبه می‌شود. این معیار می‌تواند در سطح جمله (P6) و یا تمام متن (P7) به‌دست آید.

نتایج حاصل از ارزیابی سیستم پیشنهادی و مقایسهٔ آن با سیستم‌های خلاصه‌ساز «ایجاز» و «فارسی‌سام»، در جدول (۵) آورده شده است. در روش پیشنهادی، امتیازدهی به جملات به دو صورت مورد ارزیابی قرار گرفته است. روش اول که در جدول با نام امتیازدهی اولیه آورده شده، به تشابه با جملات انتخاب‌شدهٔ قبلی توجه نمی‌شود. این در حالی است که در روش امتیازدهی ترکیبی علاوه بر امتیاز خام جملات به میزان تشابه آن‌ها به جملات انتخابی قبل نیز توجه شده است. همان‌طور که دیده می‌شود، در اغلب پارامترها سیستم پیشنهادی نتایج بهتری را به‌دست آورده است. همچنین، استفاده از شیوهٔ امتیازدهی ترکیبی به نسبت روش امتیازدهی اولیه باعث افزایش دقت در استخراج جملات خلاصه شده است.

شکل ۲، نمودار میله‌ای این مقایسه را نشان می‌دهد که در آن روش پیشنهادی با در نظر گرفتن امتیازدهی ترکیبی لحاظ شده است.

جدول ۵. نتایج ارزیابی سیستم پیشنهادی

سیستم پیشنهادی با دو روش امتیازدهی		ایجاز	فارسی‌سام	پارامترهای تابع		
امتیازدهی ترکیبی	امتیازدهی اولیه					
۰/۳۵۴۶	۰/۳۳۵۶	۰/۳۴۱۴	۰/۲۸۲۴	P1	تشابه Nتایی‌های مشابه در سطح متن	
۰/۳۶۴۶	۰/۳۴۵۲	۰/۳۳۷۸	۰/۲۹	P2	تشابه Nتایی‌های مشابه در سطح جمله	
۰/۵۲۱	۰/۵۱۲۱	۰/۳۸۹۷	۰/۳۵۳۱	P3	طولانی‌ترین زیررشتهٔ مشترک	
۰/۲۶۹۶	۰/۲۵۶۴	۰/۲۲۶۳	۰/۱۶۰۷	P4	تشابه جفت کلمات با طول آزاد	
۰/۵۱۲۱	۰/۴۹۹	۰/۴۸۶۲	۰/۴۰۱۴	P5	تعداد کلمات مشابه	
۰/۳۹۶۴	۰/۳۷	۰/۳۶۵۸	۰/۳۱۵۷	P6	معیار ویژهٔ Nتایی‌های مشترک در سطح جمله	
۰/۳۸۷۸	۰/۳۷۸۳	۰/۳۷۱۰	۰/۳۰۵۸	P7	معیار ویژهٔ Nتایی‌های مشترک در سطح متن	



شکل ۲. نمودار میله‌ای مقایسهٔ روش پیشنهادی با روش‌های «ایجاز» و «فارسی‌سام»

٦. نتیجه‌گیری

با در نظر گرفتن حجم وسیع و روزافزون اطلاعات متنی که به‌صورت الکترونیکی و در فضای وب در اختیار کاربران قرار دارد و از طرفی کمبود وقت برای بررسی تمام آن‌ها، مسئلهٔ خلاصه‌سازی متون خبری، کتاب‌ها، مقالات و ... به یکی از زمینه‌های مورد توجه محققان تبدیل شده است. در این مقاله، روشی جدید برای خلاصه‌سازی خودکار متون

فارسی ارائه گردیده که در آن با استفاده از ویژگی‌های مختلف مستخرج از متن اصلی، جملاتی با کمترین شباهت و بیشترین امتیاز برای متن خلاصه انتخاب می‌شوند.

به‌منظور ارزیابی سیستم پیشنهادی از دادگان پاسخ به همراه سیستم ارزیاب خودکار استفاده گردید. نتایج حاصل از این ارزیابی با نتایج حاصل از سیستم‌های «ایجاز» و «فارسی‌سام» مقایسه و مشخص شد در بسیاری از فاکتورهای ارزیابی، سیستم پیشنهادی موفق‌تر عمل کرده است.

فهرست منابع

Harman, Donna, and Paul Over. 2002. The DUC Summarization Evaluations. In *Proceedings of the Second International Conference on Human Language Technology Research*, 44–51. HLT '02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. http://dl.acm.org/citation.cfm?id=1289189.1289253.

Hassel, Martin, Kth Nada, and Nima Mazdak. 2004. FarsiSum - A Persian Text Summarizer. *Proceedings, Switzerland, of the Workshop on Computational Approaches to Arabic Script-Based Languages*, January, 82–84. doi:10.3115/1621804.1621826.

Hovy, Eduard, and Chin-Yew Lin. 1998. Automated Text Summarization and the SUMMARIST System. In *Proceedings of a Workshop on Held at Baltimore, Maryland: October 13-15, 1998*, 197–214. TIPSTER '98. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1119089.1119121.

Kiyoumarsi, Farshad, Esfandiar Eslami, and Asghar Tajoddin. 2010. Optimizing Text Summarization Based on Fuzzy Logic. *Iranian Journal of Fuzzy Systems* 7 (3): 15–32.

Lin, Chin-Yew. 2004. ROUGE: A Package for Automatic Evaluation of Summaries."*Microsoft Research* July. https://www.microsoft.com/en-us/research/publication/rouge-a-package-for-automatic-evaluation-of-summaries. (Accessed Jul. 15, 2015)

Luhn, H. P. 1958. The Automatic Creation of Literature Abstracts. *IBM J. Res. Dev.* 2 (2): 159–165. doi:10.1147/rd.22.0159.

Moghaddas, B. Behmadi, M. Kahani, S. A. Toosi, A. Pourmasoumi, and A. Estiri. 2013. Pasokh: A Standard Corpus for the Evaluation of Persian Text Summarizers. In Proceedings of the 3rd International eConference on Computer and Knowledge Engineering (ICCKE), 471–475. doi:10.1109/ICCKE.2013.6682873.

Nenkova, Ani, and Kathleen McKeown. 2012. A Survey of Text Summarization Techniques. In *Mining Text Data*, 43–76. Boston, MA.: Springer. doi:10.1007/978-1-4614-3223-4_3.

Pour Masoomi, Asef, Mohsen Kahani, Seyyed Ahmad Toosi, and Ahmad Estiri. 2014. Ijaz: An Operational System for Single-Document Summarization of Persian News Texts. *Signal and Data Processing* 11 (1): 33–48.

Ramezani, Majid, and Mohammad-Reza Feizi-Derakhshi. 2015. Ontology-Based Automatic Text Summarization Using FarsNet. *Advances in Computer Science : An International Journal* 4 (2): 88–96.

Shamsfard, Mehrnoush. 2011. Challenges and Open Problems in Persian Text Processing. *Proceedings of LTC* 11. http://hnk.ffzg.hr/bibl/ltc2011/book/papers/MPLRL-6.pdf. (Accessed Jul. 15, 2015)

Shamsfard, M., T. Akhavan, and M. E. Jourabchi. 2009. Parsumist: A Persian Text Summarizer. In *International Conference on Natural Language Processing and Knowledge Engineering,* NLP-KE

2009, Dalian, China: 1–7. doi:10.1109/NLPKE.2009.5313844.

Zamanifar, Azadeh, and Omid Kashefi. 2011. AZOM: A Persian Structured Text Summarizer. In *Natural Language Processing and Information Systems*, 234–37. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-22327-3_27.

Zamanifar, Azadeh, Behrouz Minaei-Bidgoli, and Mohsen Sharifi. 2008. A New Hybrid Farsi Text Summarization Technique Based on Term Co-Occurrence and Conceptual Property of the Text. In *Proceedings of the 2008 Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, 635–639. SNPD '08. Washington, DC, USA: IEEE Computer Society. doi:10.1109/SNPD.2008.57.

محمود سلطانی

متولد سال ۱۳۶۴ دارای مدرک تحصیلی کارشناسی ارشد در رشته مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه تهران است. ایشان هم‌اکنون مربی گروه مهندسی کامپیوتر دانشگاه صنعتی قوچان است.

پردازش زبان‌های طبیعی و داده‌کاوی از جمله علایق پژوهشی وی است.

جلال‌الدین نصیری

متولد سال ۱۳۶۲ دارای مدرک تحصیلی دکتری در رشته مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه تربیت مدرس است. ایشان هم‌اکنون استادیار گروه زبان‌شناسی رایانشی پژوهشکده علوم اطلاعات و مدیر آزمایشگاه متن کاوی و یادگیری ماشین در پژوهشگاه علم و فناوری اطلاعات ایران (ایرانداک) است.

پردازش زبان‌های طبیعی و یادگیری ماشین از جمله علایق پژوهشی وی است.

احسان عسکریان

متولد سال ۱۳۶۳ دارای مدرک دکتری در رشته نرم‌افزار از دانشگاه فردوسی است.

نظرکاوی، پردازش متن و داده‌کاوی از جمله علایق پژوهشی وی است.