

# ارائه سیستم خلاصه‌ساز متون فارسی بر مبنای ویژگی‌های زبان‌شناختی و رگزیون

محمود سلطانی

کارشناسی ارشد مهندسی کامپیوتر  
مریی، دانشگاه مهندسی فناوری‌های نوین قوچان

جلال الدین نصیری \*

دکتری مهندسی کامپیوتر  
استادیار پژوهشکده علوم اطلاعات

احسان عسکریان

دکتری مهندسی کامپیوتر  
پژوهشگر، دانشگاه فردوسی مشهد

دریافت: ۹۵/۱۱/۲۰ | پذیرش: ۹۶/۰۶/۲۸

فصلنامه علمی پژوهشی  
پژوهشگاه علوم و فناوری اطلاعات ایران  
شاپا(چاپی) ۸۲۲۳-۲۲۵۱  
شاپا(الکترونیکی) ۸۲۳۱-۲۲۵۱  
نمایه در SCOPUS، LISTA و ISC  
http://jipm.irandoc.ac.ir  
دوره XX | شماره X | صص XX-XX  
۱۳XX X

نوع مقاله: پژوهشی

**چکیده:** گسترش روزافزون داده‌های متنی فارسی در فضای اینترنت و پیچیدگی جستجو 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 به صورت 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 بدست می‌آید. وزن بهینه هر ویژگی در این ترکیب از روش رگزیون خطی و با استفاده از پیکره آموزشی پاسخ محاسبه شده است. پس از محاسبه 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، دارای کمترین میزان شباهت با جملات انتخاب‌شده در مراحل قبلی نیز است. نتایج بدست آمده از مقایسه الگوریتم پیشنهادی با دو سیستم خلاصه‌ساز ایجاز و فارسی‌سام با استفاده از پیکره پاسخ، نشان می‌دهد در بیشتر معیارهای ارزیابی پیشرفت قابل توجهی حاصل شده است.

**کلیدواژه‌ها:** خلاصه‌سازی تک‌سندی، ویژگی‌های زبان‌شناختی متن، رگزیون خطی

\*پدیدآور رابط J.Nasiri@Irandoc.ac.ir

به این مقاله به شکل زیر استناد کنید:

دورن متن:

(سلطانی، نصیری، عسکریان، زودآیند)

در فهرست منابع:

سلطانی، محمود، نصیری، جلال، عسکریان، احسان. زودآیند. ارائه سیستم خلاصه‌ساز متون فارسی بر مبنای ویژگی‌های زبان‌شناختی و رگزیون. پژوهشنامه پردازش و مدیریت اطلاعات.

(دسترسی در http://jipm.irandoc.ac.ir//

روز/ماه/سال)

۱. مقدمه

رشد سریع وب و منابع داده موجود در آن باعث شده است 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 جستجو کمک کند. ایده اصلی در خلاصه‌سازی خودکار پیدا کردن یک نسخه مختصرتر از سند اصلی توسط ماشین است به طوری 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 خلاصه اخبار، بازیابی اطلاعات و ... کاربرد داشته باشد.

روش‌های خلاصه‌سازی خودکار متون را می‌توان به دو دسته استخراجی<sup>۱</sup> و چکیده‌ای<sup>۲</sup> تقسیم کرد. در خلاصه‌سازی استخراجی، ساختار جملات متن اصلی تغییر نمی‌کند. در این روش - ها جملات متن اصلی بر اساس ویژگی‌های مشخص وزن دهی شده و جملات با وزن بیشتر به عنوان جملات مهم تر در خلاصه 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 دسته تک‌سندی<sup>۳</sup> و چند سندی<sup>۴</sup> تقسیم‌بندی می‌شود، به طوری که در خلاصه‌سازی تک‌سندی، یک سند به عنوان سند اولیه خلاصه می‌شود و در خلاصه‌سازی چند سندی، خلاصه‌ای از چندین سند مرتبط از لحاظ موضوعی تهیه می‌شود و هدف تهیه یک نسخه خلاصه از ترکیب این اسناد است. پیچیدگی‌ها در روش‌های چند سندی بیشتر از روش‌های تک‌سندی است چرا که ممکن است چندین سند با ارتباط موضوعی اما با دیدگاه‌های متفاوت و حتی متناقض داشته باشیم.

روش‌های اولیه خلاصه‌سازی متون از ویژگی‌های ساده آماری مانند تعداد تکرار کلمات، طول و موقعیت قرار گرفتن جمله در متن استفاده می‌کردند. این در حالی است که در سال‌های اخیر استفاده از ابزارهای جدید در زمینه پردازش متن، بکارگیری روش‌های یادگیری ماشین و همچنین تحلیل‌های معنایی توانسته است تا حدود زیادی کیفیت سیستم‌های خلاصه‌ساز را افزایش

<sup>1</sup> Extractive

<sup>2</sup> Abstractive

<sup>3</sup> Single- Document summarization

<sup>4</sup> Multi-Document summarization

دهد. در اغلب سیستم‌های خلاصه‌ساز 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، درخت تصمیم، مدل مخفی مارکوف، شبکه‌های عصبی و ... استفاده شده است. با توجه به تعداد زیاد روش‌های پیشنهادی به منظور خلاصه‌سازی خودکار متون باید روشی برای ارزیابی و مقایسه این روش‌ها وجود داشته باشد. در سال‌های گذشته تلاش‌های بسیاری برای این منظور صورت گرفته است. برای نمونه می‌توان به انجمن (DUC) (Harman and Over 2002) اشاره کرد که به منظور ارزشیابی سیستم‌های خلاصه‌ساز هر ساله با در اختیار قرار دادن پیکره‌ها و دادگان مناسب به برگزار همایش و رقابت می‌پردازد. در این میان معیارهای مختلفی برای ارزشیابی و مقایسه الگوریتم‌های خلاصه‌ساز ارائه شده است. در این زمینه ROUGE (Lin 2004) مجموعه‌ای از معیارها و نرم‌افزارهایی است که برای ارزشیابی سیستم‌های خلاصه‌ساز خودکار و ترجمه ماشینی استفاده می‌شود. متاسفانه برای زبان فارسی پیکره‌های محدودی برای ارزشیابی و یادگیری سیستم‌های خلاصه‌ساز وجود دارد. پیکره پاسخ (Moghaddas et al. 2013) یکی از این نمونه‌ها است که دارای ۱۰۰ متن از اخبار با موضوعات مختلف به همراه پنج خلاصه انسانی برای هر کدام است. در این مقاله از پیکره پاسخ به منظور یادگیری سیستم و ارزشیابی نهایی آن استفاده شده است.

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

## ۲. معرفی مسئله

مسئله‌ای که در این مقاله به آن پرداخته شده است، مسئله خلاصه‌سازی خودکار متون خبری است. خلاصه‌سازی خودکار به فرایند تجمیع و فشرده‌سازی یک یا چند منبع اطلاعاتی به منظور کوتاه کردن متن و یا پاسخگویی به درخواست کاربر بدون دخالت انسان گفته می‌شود. همانطور که در این تعریف آمده است، خلاصه‌سازی از جنبه‌های مختلف قابل بررسی و دسته‌بندی است. در (Nenkova and McKeown 2012) به بررسی این زوایا پرداخته شده است که در ادامه به برخی از آنها اشاره می‌شود.

یکی از موضوعات قابل بررسی هدف از تهیه نسخه خلاصه است. بر این اساس تهیه یک خلاصه می‌تواند به منظور پاسخگویی به پرس و جویی از سمت کاربر باشد که باید تا حد امکان خلاصه نهایی به این درخواست نزدیکتر باشد. علاوه بر این خلاصه سازی می‌تواند در راستای روزرسانی پایگاه داده و افزودن جمع‌بندی متون دیگر در قالب یک متن خلاصه باشد. جنبه قابل بررسی دیگر تعداد منابعی است که باید بر اساس آنها متن خلاصه تهیه شود. خلاصه سازی یا بر روی یک متن ورودی صورت می‌پذیرد، که به آن خلاصه سازی تک سندی گفته می‌شود و یا باید یک نسخه خلاصه از چندین متن موجود تهیه شود، که به آن خلاصه سازی چند سندی گفته می‌شود. روش تهیه خلاصه یکی دیگر از جنبه‌های تفاوت سیستم‌های خلاصه ساز است. خلاصه سازی می‌تواند بصورت استخراجی باشد که در آن متن خلاصه زیر بخشی از متن اصلی است. به عبارت دیگر گزیده‌ای از متن اصلی بدون تغییر ساختاری در متن خلاصه ظاهر می‌شود. نوع دیگری که به مراتب چالش برانگیزتر از نوع قبلی است، خلاصه سازی چکیده‌ای است. در این نوع خلاصه سازی، جملاتی با ساختار جدید توسط ماشین و برگرفته از متن اصلی ساخته می‌شود که علاوه بر کوتاه‌تر بودن نسبت به متن اصلی، مفهوم کلی آنرا بیان می‌کند. همان طور که دیده می‌شود در این نوع خلاصه سازی جنبه‌های دستوری و نگارشی زبان مورد نظر باید در نظر گرفته شود.

با توجه به جنبه‌های مختلف خلاصه سازی که در بالا بیان گردید، در این مقاله، تمرکز به روی خلاصه سازی تک سندی به روش استخراجی است.

### ۳. کارهای مرتبط

اولین تلاش‌ها به منظور خلاصه‌سازی خودکار متون، به کار انجام شده در (Luhn 1958) بر می‌گردد که 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 است که بسیاری از آنها بصورت تک سندی و استخراجی هستند. به عنوان مثال در (Hovy and Lin 1998) روشی بر اساس امتیازدهی به جملات و انتخاب جملات مهم ارائه شده است. در ادامه با توجه به فراوانی تعداد روش‌های خلاصه سازی متون فقط به بررسی برخی از روش‌ها در زبان فارسی پرداخته شده است.

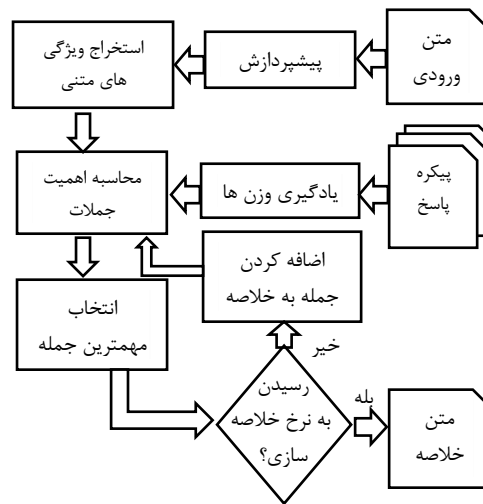
در سال‌های اخیر تلاش‌های زیادی به منظور خلاصه‌سازی خودکار متون فارسی انجام شده است (M. Shamsfard, Akhavan, and Jourabchi 2009; Kiyoumars, Eslami, and Tajoddin 2010; Zamanifar, Minaei-Bidgoli, and Sharifi 2008; Hassel, Nada, and Mazdak 2004; Pour Masoomi et al. 2014) و.... در (M. Shamsfard, Akhavan, and Jourabchi 2009) در این سیستم‌ها

چندین ویژگی متنی به منظور امتیازدهی به جملات مورد استفاده قرار گرفته است. در هر مرحله از الگوریتم جمله‌ای که به جملات انتخاب شده در مراحل قبل بیشترین شباهت را دارد حذف می‌شود و از جملات باقی مانده جمله با اهمیت بیشتر انتخاب می‌شود. در (Kiyoumars, Eslami, and Tajoddin 2010) روشی ارائه شده است که با استفاده از وردنت و منطق فازی مناسب‌ترین جملات را به عنوان جملات خلاصه نهایی انتخاب می‌کند. روش ارائه شده در (Zamanifar, Minaei- Bidgoli, and Sharifi 2008) با استفاده از هم رخداد کلمات برای پیدا کردن کلمات مهم و همچنین با استفاده از مترادف کلمات برای جلوگیری از انتخاب جملات مشابه، به خلاصه سازی متون می‌پردازد که نتایج آن به نسبت (Hassel, Nada, and Mazdak 2004) بهتر گزارش شده است. الگوریتمی که در (Hassel, Nada, and Mazdak 2004) پیشنهاد شده است تنها بر اساس اطلاعات آماری مستخرج از متن است که در واقع نسخه فارسی از الگوریتم SwuSum است. این الگوریتم برای خلاصه‌سازی متون زبان سوئدی مورد استفاده قرار می‌گیرد. ایجاز (Pour Masoomi et al. 2014) یکی دیگر از سیستم‌های خلاصه‌ساز خودکار برای زبان فارسی است که در آن از نه ویژگی متنی به منظور ارزش گذاری جملات استفاده شده است. با استفاده از روش کمترین مربعات وزن هر ویژگی محاسبه شده و در نهایت جملات با بیشترین درجه اهمیت انتخاب می‌شوند. این سیستم بصورت تک سندی و چند سندی می‌تواند خلاصه مناسب را تولید نماید. در (Zamanifar and Kashefi 2011) از ویژگی‌های آماری و مفهومی متن استفاده کرده و پس از تبدیل ساختار متن ورودی به درخت فراکتال، خلاصه‌ای از آن تولید می‌کند. (Ramezani and Feizi-Derakhshi 2015) یکی دیگر از روش‌هایی است که در آن از هستان‌شناسی استفاده شده است. در الگوریتم پیشنهادی از مجموعه واژگان فارسی نت<sup>1</sup> به منظور پیدا کردن رابطه معنایی بین بخش‌های مختلف متن استفاده شده است. این روابط در یک گراف مبتنی بر متن پیاده سازی شده و در نهایت از روی آن جملات مهم برای خلاصه نهایی انتخاب می‌شوند. همانطور که دیده می‌شود شیوه کاری در اغلب روش‌های بررسی شده، خلاصه سازی استخراجی و تک سندی است. تعدادی از آن‌ها بر پایه اطلاعات آماری و ساختاری متن و تعدادی بر اساس اطلاعات مفهومی و هستان‌شناسی بنا نهاده شده‌اند. علاوه بر این فرایند انتخاب جملات مهم در آنها متفاوت است. روش‌های مبتنی بر نظریه گراف‌ها، منطق فازی، دسته بندی و خوشه بندی نمونه‌هایی از روش‌های مورد استفاده است.

#### ۴. روش پیشنهادی

<sup>1</sup> FarsNet

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



شکل ۱: مراحل انجام الگوریتم پیشنهادی

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

#### ۴-۱ پیش‌پردازش متن

به علت وجود برخی چالش‌های نگارشی و دستوری در زبان فارسی (Mehrnoush Shamsfard, 2011)، انجام برخی پیش‌پردازش‌ها بر روی متن اولیه برای رسیدن به دقت بیشتر خلاصه‌سازی لازم و ضروری است. در این فاز چهار اقدام اصلی وجود دارد: جداسازی جملات، جداسازی کلمات، حذف کلمات توقف ۱ و ریشه‌یابی کلمات. در مرحله جداسازی، با استفاده از علائم جداکننده (،، ؟، و ...) و یک سری قواعد خاص جملات متن ورودی استخراج می‌شود. در ادامه با توجه به

<sup>1</sup> Stop List

ویژگی‌های دستوری زبان فارسی، کلمات هر جمله جدا شده و پس از آن کلمات پرتکرار و بی-اهمیت که کلمات توقف نامیده می‌شوند (مانند: از، که، است و ...) از لیست کلمات حذف می‌شوند. ریشه‌یابی کلمات آخرین مرحله از پیش پردازش است که در آن ریشه کلمات با کلمه اصلی جایگزین می‌شود. این کار به منظور افزایش دقت در تحلیل‌های آماری و بسامد کلمات است.

#### ۴-۲ استخراج ویژگی‌های متنی

در خلاصه‌سازی استخراجی، مینا انتخاب مهمترین جملات متن برای خلاصه نهایی است. به منظور 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 پرکاربرد 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 استخراج شده برای هر جمله

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

#### ۴-۲-۱ تشابه با عنوان

استفاده از عنوان یکی از ابتدایی‌ترین شاخص‌ها برای انتخاب جملات متن خلاصه است. میزان شبهت هر جمله با عنوان متن بر اساس روش‌های سنجش میزان شبهت 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 محاسبه اهمیت یک جمله در این مقاله استفاده شده است. روش محاسبه این معیار در فرمول (۱) آورده شده است.

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

$$F1_i = \frac{\text{sim}(T, S_i)}{\text{Max}_k(\text{sim}(T, S_k))} \quad (1)$$

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

#### ۴-۲-۲ طول جملات

جملات خیلی کوتاه و یا خیلی طولانی معمولاً جملات مفیدی برای متن خلاصه نیستند. بطور کلی در جملات کوتاه اطلاعات مفیدی دیده نمی شود و جملات طولانی با توجه به ذات خلاصه سازی نمی توانند انتخاب مناسبی باشند. می توان با تعریف یک حد آستانه برای طول جمله، جملاتی که در این محدوده قرار نگیرند را حذف کرد.

البته در روش پیشنهادی هیچ جمله ای به این علت حذف نمی شود و تنها این معیار مانند معیارهای دیگر در محاسبه ارزش جملات بکار گرفته می شود. به عبارت دیگر تصمیم گیری برای انتخاب و یا عدم انتخاب یک جمله به تمام ویژگی های آن بستگی دارد و ممکن است ویژگی های دیگر، اهمیت جمله ای با طول زیاد و یا طول کم را بالا برده و الگوریتم آن جمله را برای متن خلاصه انتخاب کند. نحوه محاسبه این معیار در فرمول (۲) آورده شده است.

$$F2_i = \frac{|S_i|}{\text{Max}_k(|S_k|)} \quad (2)$$

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

#### ۴-۲-۳ شباهت با زمینه متن

جملاتی که به زمینه اصلی متن شبیه ترند، می توانند گزینه مناسب تری برای انتخاب در خلاصه نهایی باشند. زمینه اصلی متن را می توان از روی کلمات پرتکرار به دست آورد. کلمات پرتکرار در متن در واقع مشخص کننده زمینه متن هستند. بنابراین جملاتی که شامل تعداد بیشتری از این کلمات پرتکرار باشند، به زمینه متن نزدیک تر هستند. فرمول (۳) به منظور محاسبه شباهت جملات با زمینه متن اصلی استفاده شده است.

$$F3_i = \frac{\sum_{j=1}^l \text{tf}(w_j^i)}{\text{Max}_k(\sum_{j=1}^l w_j^k)} \quad (3)$$

که در آن  $\text{tf}(w_j^i)$  بسامد کلمه  $z$  در جمله  $i$  و  $l$  طول جمله است.

#### ۴-۲-۴ موقعیت جمله



جایگاه جملات در متن تا حدود بسیار زیادی می تواند روی اهمیت آنها تأثیر گذار 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. هر چند که در برخی موارد به علت عدم رعایت اصول نگارشی این موضوع تحت تأثیر قرار می گیرد. با این حال در این مقاله از فرمول (۴) به منظور محاسبه تأثیر جایگاه جملات استفاده شده است.

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

که در آن  $\text{pos}(S_i)$  موقعیت جمله  $i$ ام در متن است.

#### ۴-۲-۵ عبارات خاص

وجود برخی عبارات خاص در جملات اهمیت ویژه ای به جمله می دهد. در مقابل عباراتی هستند که وجود آنها در جمله از اهمیت آن کم می کنند. برای محاسبه نقش این عبارات در اهمیت جمله، از پنج نوع عبارت خاص استفاده شده است: عبارات مهم، عبارات غیر مهم، القاب، ضمائر و علائم نقل قول. به عنوان مثال جمله ای که شامل عبارت «در نتیجه» باشد به احتمال زیاد حاوی اطلاعات مفید است و در مقابل جمله ای که در آن از عبارت «به عنوان مثال» استفاده شده است نمی تواند گزینه خوبی برای متن خلاصه باشد. جدول ۲ شامل برخی از این عبارات خاص می باشد. در فرمول (۵) نحوه محاسبه تأثیر عبارات خاص در اهمیت جمله آورده شده است.

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

که در آن  $\text{im}(S_i)$  تعداد عبارات مهم،  $h(S_i)$  تعداد القاب،  $\text{uim}(S_i)$  تعداد عبارات غیر مهم،  $\text{pr}(S_i)$  تعداد ضمائر و  $\text{qu}(S_i)$  تعداد علائم نقل قول جمله است.

جدول ۲: مثال هایی از عبارات خاص در جملات

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

#### ۴-۲-۶ اسامی خاص

معمولا وجود اسامی خاص در جمله می تواند به اهمیت آن بیافزاید. بر این اساس جمله ای 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 خلاصه سازی داشته باشند. فرمول (۶) برای محاسبه مقدار این ویژگی استفاده می‌شود.

$$F6_i = \frac{NE(S_i)}{\text{Max}_k(NE(S_k))} \quad (۶)$$

که در آن  $NE(S_i)$  تعداد اسامی خاص در جمله  $S$  می‌باشد.

#### ۴-۲-۷ اطلاعات عددی

معمولا جملاتی که دارای اطلاعات عددی مانند تاریخ، درصدها و ... از اهمیت بیشتری برخوردار هستند. فرمول (۷) نرمال شده تعداد اطلاعات عددی در جمله است.

$$F7_i = \frac{ND(S_i)}{\text{Max}_k(ND(S_k))} \quad (۷)$$

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

#### ۴-۳ محاسبه میزان اهمیت هر جمله

در این مرحله برای هر جمله هفت مقدار بر اساس پارامترهای توضیح داده شده در بخش قبل بدست آمده است که باید بر اساس آنها میزان اهمیت هر جمله را محاسبه کرد. نقش هر پارامتر در تعیین اهمیت جمله یکسان نیست؛ به عبارت دیگر ترکیب پارامترها با وزن یکسان بخوبی گویای اهمیت جملات نیست. برای محاسبه وزن هر پارامتر در تعیین اهمیت جمله می‌توان از روش‌های یادگیری ماشین استفاده نمود.

یکی از این روش‌ها رگرسیون خطی است. ما از نرم افزار WEKA برای یادگیری وزن‌ها بر اساس این روش استفاده کرده‌ایم. برای این منظور از ۲۰۰۰ جمله پیکره پاسخ استفاده شده است. با توجه به اینکه در پیکره پاسخ هر متن خلاصه دارای پنج خلاصه انسانی است به هر جمله بسته به اینکه در چند خلاصه انسانی ظاهر شده است یک عدد بین صفر تا پنج انتساب داده می‌شود (C)؛ بنابراین به جمله ای که در هیچ خلاصه انسانی وجود ندارد عدد صفر و به جمله ای که در تمام خلاصه‌های انسانی یک متن دیده شده است عدد پنج تخصیص داده می‌شود.

جدول ۳ بخشی از دادگان یادگیری را نشان می‌دهد که برای هر جمله مقدار هفت ویژگی و عدد انتساب داده شده به آن (C) آورده شده است.

جدول ۳: بخشی از دادگان آموزشی

شماره جمله	F1	F2	F3	F4	F5	F6	F7	تعداد خلاصه 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

#### ۴-۴ انتخاب جملات متن خلاصه

هدف اصلی در خلاصه سازی استخراجی انتخاب جملات مهم است. یکی از روش های بدست آوردن جملات مهم، انتساب یک مقدار عددی به هر جمله به عنوان میزان اهمیت و انتخاب جملات با بیشترین این مقدار است. ما در روش پیشنهادی برای بدست آوردن میزان اهمیت هر جمله از ترکیب هفت ویژگی استخراج شده از متن استفاده کرده ایم. این هفت ویژگی و نحوه استخراج آنها در قسمت های قبل توضیح داده شد. در این بخش نحوه محاسبه میزان اهمیت جملات از روی این ویژگی شرح داده می شود.

پس از پیش پردازش های لازم بر روی متن ورودی، مقادیر هفت ویژگی برای هر جمله استخراج می شود. پس از آن و بر اساس وزن هر ویژگی که با استفاده از روش رگرسیون خطی بدست آمده است، با استفاده از فرمول (۸) می توان میزان اهمیت جملات را محاسبه کرد. در جدول (۴) این شیوه امتیازدهی، روش اولیه نامیده شده است.

$$\text{Score}(S_i) = \frac{\sum_{k=0}^n F_k * W_k}{n} \quad (8)$$

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

یکی از نقاط ضعف این شیوه برای محاسبه میزان اهمیت جملات عدم توجه به میزان شباهت جملات انتخابی در خلاصه نهایی است. با این شیوه امتیازدهی، در صورتی که دو جمله ی شبیه به هم در متن اصلی، امتیاز بالایی کسب کنند، هر دو جمله برای متن خلاصه انتخاب می شوند. این انتخاب باعث ایجاد افزونگی در متن خلاصه می شود. برای جلوگیری از این مشکل روش امتیازدهی جدیدی بر اساس بیشترین شباهت حاشیه ای ارائه شده است.

بیشترین شباهت حاشیه ای: در اغلب سیستم های بازیابی اطلاعات، لیستی از اسناد مرتبط با درخواست کاربر آماده می شود که بر اساس میزان تشابه سند با درخواست کاربر و همچنین کمترین شباهت با اسناد قبلی مرتب می شود. ترکیب خطی این دو فاکتور شباهت حاشیه ای نامیده

(۹)

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

می‌شود؛ بنابراین سندی با شباهت حاشیه‌ای بیشتر رتبه بالاتری در لیست اسناد بازیابی شده دارد. فرمول (۹) نحوه محاسبه شباهت حاشیه‌ای را نشان می‌دهد.

که در آن  $c$  مجموعه کل اسناد،  $Q$  درخواست کاربر،  $R$  لیست امتیاز دهی شده از اسناد بازیابی شده و  $S$  زیرمجموعه  $R$  از 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 محاسبه شده برای آن از فرمول (۸) بیشترین بوده، دوماً دارای کمترین شباهت با جملات انتخاب شده قبلی باشد. فرمول (۱۰) بیان‌کننده نحوه محاسبه امتیاز هر جمله بر اساس شباهت حاشیه‌ای است. در جدول (۴) این شیوه امتیازدهی، روش ترکیبی نامیده شده است.

$$New\_Score(S_i \in Q) = \lambda * Score(S_i) - (1 - \lambda) * Max_k(sim(S_i, R_k)) \quad (10)$$

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

برای بدست آوردن مناسب‌ترین مقدار  $\lambda$  الگوریتم را بر مبنای مقادیر متفاوت  $\lambda$  و بر روی دادگان آزمایشی پاسخ اجرا کرده و بر اساس نتایج حاصل شده بهترین مقدار  $\lambda = 0.6$  انتخاب گردید. جدول (۴) نتایج حاصل را نشان می‌دهد. همانطور که دیده می‌شود، مقدار  $\lambda = 0.6$  در اقل 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 فشرده سازی برسیم؛ به عبارت دیگر تعداد کلمات متن خلاصه به درصد از پیش تعیین شده‌ای از تعداد کل کلمات متن اصلی برسد. فرمول (۱۱) نحوه محاسبه نرخ فشرده سازی را نشان می‌دهد.

#### جدول ۴: نتایج اجرا الگوریتم برای مقادیر مختلف $\lambda$

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

$$\text{Compressionrate} = \frac{\text{lengthofsummary}}{\text{lengthoforiginal}} \quad (11)$$

در اولین مرحله از الگوریتم با توجه به اینکه در فرمول (۹) مجموعه R شامل جمله‌ای نیست، مبنای محاسبه‌ی امتیاز جملات فرمول (۸) است. در مراحل بعدی ابتدا بر اساس فرمول (۹) امتیاز جملات باقیمانده محاسبه مجدد شده و جمله‌ای که بیشتر امتیاز را کسب کرده به لیست جملات متن خلاصه افزوده می‌شود. در ادامه الگوریتم با توجه به فرمول (۱۱) نرخ فشرده سازی تا این مرحله را محاسبه کرده و با مقدار از پیش تعیین شده قبلی مقایسه می‌کند. در صورتی که نرخ فشرده سازی از این مقدار کمتر باشد و یا به عبارت دیگر هنوز خلاصه نهایی کامل نشده باشد، الگوریتم ادامه پیدا کرده و پس از محاسبه مجدد امتیاز جملات انتخاب نشده، مراحل تا رسیدن به نرخ فشرده سازی 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 خلاصه ساز زبان فارسی (ایجاز و فارسی سام) مقایسه شده است. علاوه بر این پیکره مورد استفاده برای ارزیابی پیکره پاسخ (Moghaddas et al. 2013) انتخاب شده است. پاسخ اولین پیکره متنی برای ارزیابی سیستم‌های خلاصه ساز فارسی است که در دو بخش تک سندی و چند سندی تهیه شده است. در بخش تک سندی، یکصد متن از اخبار در شش حوزه مختلف (اقتصادی، اجتماعی، فرهنگی، ورزشی، سیاسی و علمی) به همراه پنج متن خلاصه برای هر کدام وجود دارد.

در ارزیابی خلاصه سازی از چندین معیار استفاده می‌شود که در ادامه به بررسی برخی از آنها که در این مقاله استفاده شده، پرداخته شده است.

#### ۱-۵ N تایی‌های مشترک

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

#### ۲-۵ طولانی‌ترین زیر رشته 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.

#### ۳-۵ تعداد جفت کلمات مشترک با فاصله آزاد

به هر جفت کلمه (با حفظ ترتیب) در جمله skip-gram گفته می‌شود. این معیار با اندازه گیری تعداد skip-gram های مشترک بین خلاصه ماشینی و خلاصه مرجع محاسبه می‌شود. این معیار با عنوان P4 در جداول نمایش داده شده است.

#### ۴-۵ تعداد کلمات مشترک

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

#### ۵-۵ معیار ویژه N تایی‌های مشترک

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

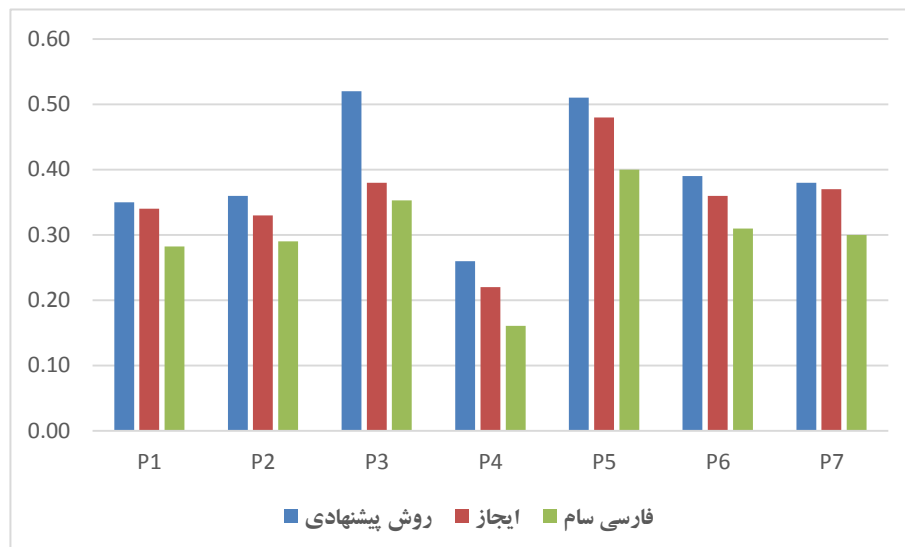
نتایج حاصل از ارزیابی سیستم پیشنهادی و مقایسه آن با سیستم‌های خلاصه ساز ایجاز و فارسی‌سام، در جدول (۵) آورده شده است. در روش پیشنهادی امتیاز دهی به جملات به دو صورت مورد ارزیابی قرار گرفته است. روش اول که در جدول با نام امتیازدهی اولیه آورده شده است، به تشابه با جملات انتخاب شده قبلی توجه نمی‌شود. این در حالی است که در روش امتیاز-دهی 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 بدست آورده است. همچنین استفاده از شیوه 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 امتیاز دهی ترکیبی لحاظ شده است.

جدول (۵): نتایج ارزیابی سیستم پیشنهادی

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



شکل ۲: نمودار میله‌ای مقایسه روش پیشنهادی با روش‌های ایجاز و فارسی-سام

## ۶. نتیجه گیری

با در نظر گرفتن حجم وسیع و روزافزون اطلاعات متنی که بصورت الکترونیکی و در فضای وب در اختیار کاربران قرار دارد و از طرفی کمبود وقت برای بررسی تمام آنها، مسئله خلاصه سازی متون خبری، کتاب‌ها، مقالات و ... به یکی از زمینه‌های مورد توجه محققان تبدیل شده

است. در این مقاله، روشی جدید برای خلاصه سازی 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.

به منظور 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 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.
- Kiyoumars, 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/>.
- 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 *ICCKE 2013*, 471–75. doi:10.1109/ICCKE.2013.6682873.
- Nenkova, Ani, and Kathleen McKeown. 2012. "A Survey of Text Summarization Techniques." In *Mining Text Data*, 43–76. Springer, Boston, MA. 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, M., T. Akhavan, and M. E. Jourabchi. 2009. "Parsumist: A Persian Text Summarizer." In *International Conference on Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009*, 1–7. doi:10.1109/NLPKE.2009.5313844.
- Shamsfard, Mehrnosh. 2011. "Challenges and Open Problems in Persian Text Processing." *Proceedings of LTC 11*. <http://hnk.ffzg.hr/bibl/lc2011/book/papers/MPLRL-6.pdf>.
- 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. Springer, Berlin, Heidelberg. 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.

## **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 Advanced Technology  
m.soltani@qiet.ac.ir

### **Jalal Nasiri**

PhD in Computer Engineering  
Assistant Professor, Iranian Research Institute for Information  
Science and Technology (IRANDOC)  
j.nasiri@irandoc.ac.ir

### **Ehsan Asgarian**

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

### **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**



ارائه سیستم خلاصه‌ساز متون فارسی بر مبنای ویژگی‌های زبان شناختی و رگرسیون | نصیری و دیگران

Single- Document summarization, Linguistic Feature, linear regression