نشریه مطالعات مهندسی کامپیوتر آپادانا
دوره ۱، شماره ۱، زمستان ۱۴۰۲، ۶-۱۲
ajce.apadana.ac.ir

خلاصه‌سازی انتزاعی متن

افشین سخاوتی *
دانشجوی کاردانی پیوسته مهندسی نرم افزار، مؤسسه آموزش عالی آپادانا، شیراز، ایران.

تاریخ دریافت: ۱۴۰۲/۱۰/۰۲
تاریخ پذیرش: ۱۴۰۲/۱۱/۳۰

چکیده

خلاصه‌سازی انتزاعی متن (ATS) وظیفه جمع‌آوری اطلاعات از مقالات و منابع مختلف و فشرده‌سازی آن‌ها به‌گونه ای که محتوا گم نشده و به خوبی نمایش داده شود و اطلاعاتی از دست نرود. ATS به خلاصه‌های تولید شده توسط انسان نزدیک‌تر است و قابلیت نمایش و ترکیب چندین اطلاعات را دارد. با ظهور معماری‌های عمیق (deep learning)، بسیاری از وظایف مربوط به پردازش زبان طبیعی به کارایی بالا و پایدار و قابل مقایسه‌ای دست یافتند؛ در ترجمه ماشینی، تشخیص گفتار، شرح تصاویر و بسیاری دیگر از مدل‌های توالی به ترتیب، سودمند بوده و نتایج امیدوار کننده‌ای را نشان داده است. ابزارهای زبانی مانند برچسب‌گذاری‌های بخش گفتار، Named Enity Recognizer برای زبان‌های هندی چندان رقابتی نیستند و ازاین‌رو، تکنیک‌های خاص زبان برای زبان‌های هندی، خیلی خوب عمل نمی‌کنند. تکنیک‌های عمیق زبان، اگوستیک هستند؛ ازاین‌رو می‌توانند بر این کاستی‌ها غلبه کنند. در این مقاله، شبکه‌های مولد برای ایجاد جوهره برای ایجاد قطعه طولانی‌تر متن در ارتباط با تشخیص بازنویسی تأکید می‌شوند.

در این مقاله، ما یک فرایند متضاد را برای خلاصه‌سازی انتزاعی متن پیشنهاد می‌کنیم ،که در آن ما به‌طور هم‌زمان یک مدل تولیدی G و یک مدل افتراقی D[1] آموزش می‌دهیم. به‌ویژه، ما مولد G را به‌عنوان عامل یادگیری تقویتی می‌سازیم که متن خام را به‌عنوان ورودی و خلاصه انتزاعی را پیش‌بینی می‌کند. ما همچنین یک تمایز ایجاد می‌کنیم که سعی می‌کند خلاصه تولید شده را از خلاصه حقیقت پایه متمایز کند. در این مقاله یک رویکرد ترکیبی برای تولید خلاصه انتزاعی خلاصه‌شده است که در آن جملات با استفاده از روابط سطح جمله بین جملات در ارتباط با اصل خوشه‌بندی مارکوف ، خوشه‌بندی می‌شود. سپس رتبه‌بندی جملات در هر خوشه انجام می‌شود و در صورت امکان جملات وزنی بالای هر خوشه با استفاده از برخی قواعد زبانی با بهترین عناصر آن ترکیب می‌شود.

کلیدواژه‌ها: خلاصه‌سازی انتزاعی متن**(ATS)** ، جمع‌آوری اطلاعات، فشرده‌سازی اطلاعات، ابزارهای زبانی، شبکه‌های مولد، تشخیص بازنویسی، یادگیری تقویتی

* نویسنده مسئول:afshin.se@stu.apadana.ac.ir

معرفی

خلاصه‌سازی، فرایند فشرده‌سازی یک متن منبع به نسخه کوتاه‌تر با حفظ محتوای اطلاعاتی آن است. با رشد تکاملی رسانه‌های اجتماعی، تجزیه‌وتحلیل اطلاعات موجود در آن مهم شده است. در زمان‌های اخیر، خلاصه‌سازی خودکار متن مورد توجه حوزه‌های تحقیقاتی پردازش زبان طبیعی، متن‌کاوی و بسیاری از زمینه‌های دیگر قرار گرفته است. چندین رویکرد سنتی مبتنی‌بر یادگیری ماشین، علاوه‌بر چند رویکرد مبتنی‌بر یادگیری عمیق، این حوزه را در گذشته بررسی کرده اند. در این مقاله، یک رویکرد مبتنی‌بر یادگیری عمیق مورد بررسی قرار گرفته است که از یک مدل تولیدی برای ایجاد خلاصه‌هایی از مجموعه داده‌های ورودی استفاده می‌کند. قبلاً از شبکه‌های متخاصم مولد برای تولید عنوان استفاده می‌شد (Bhargava et al., 2020).

توسعه دایرةالمعارف از خلاصه‌سازی خودکار متن با روش TextRank استفاده می‌کنند؛ زیرا از یک پرس‌وجو که جست‌وجو می‌شود، معانی زیادی از کلمات وجود دارد که باید خلاصه شوند؛ این روش با انتخاب اسناد مرتبط با پرس‌وجو شروع می‌شود، سپس اسناد انتخاب شده را با استفاده از روش TextRank خلاصه می‌کند تا براساس تمامی معانی کلمات، خلاصه‌ای به دست آورد، در نهایت نتیجه خلاصه توسط سیستم با خلاصه‌ای که به‌صورت دستی توسط انسان ساخته شده و اهداف خلاصه آزمایش می‌شود (Fakhrezi et al., 2021). خلاصه‌نویسی متن را می‌توان به‌طور کلی براساس فرم، به دو دسته تقسیم کرد که یکی استخراجی و دیگری خلاصه‌سازی انتزاعی است. کارهای زیادی در خلاصه‌سازی استخراجی از آغاز آن انجام شده است که از تلاش زبانی کمتری استفاده می‌کند، درحالی‌که مورد دوم، هنوز یک حوزه فعال تحقیقاتی است که به تلاش بیشتر برای پردازش زبان نیاز دارد. بعلاوه این را می‌توان براساس بُعد به خلاصه‌سازی تک‌سندی یا چندسندی تقسیم کرد؛ این مقاله، فقط بر خلاصه‌سازی انتزاعی یک سند متمرکز است. در خلاصه انتزاعی، به‌طور کلی سه مرحله اساسی دنبال می‌شود. در مرحله اول، مضامین مختلف صحبت شده در سند را شناسایی کنید و از هر موضوع، جملات بسیار مرتبط را استخراج کنید که می‌توان آنها را (در صورت امکان) برای تولید جمله جدید، ترکیب کرد. در مرحله دوم، کلمات غیرضروری بدون مذاکره در مورد معنای جمله از جمله حذف می‌شوند. در مرحله آخر در صورت امکان چند کلمه جدید معرفی می‌شود. در این مقاله، جملات با استفاده از ویژگی‌های معنایی و همچنین ویژگی‌های آماری، خوشه‌بندی می‌شوند تا جملات بسیار مرتبطی را که از آن جمله رتبه‌بندی شده برتر و جمله مناسب آن (اگر در خوشه یافت شود) به دست می‌آید که می‌توانند برای تولید جمله جدید ترکیب شوند، سپس از هر خوشه جمله ادغام شده (در صورت ادغام) در غیر این صورت دو جمله برتر رتبه‌بندی شده به ماژول فشرده‌سازی جمله می‌رود تا کلمات غیرضروری از جمله را برای تولید خلاصه حذف کند. مقاله به شرح زیر سازماندهی شده است: بخش ۲ کار مرتبط را ارائه می‌دهد. بخش ۳ خط لوله کامل سیستم را شرح می‌دهد. در بخش ۴ ارزیابی سیستم و سپس نتیجه‌گیری مورد بحث قرار می‌گیرد (Sahoo et al., 2018).

پیشینه تحقیق

بارگاوا و همکاران[1] (۲۰۲۰)، در این بخش، در مورد تکنیک‌های خلاصه‌سازی استخراجی و انتزاعی اخیر توضیح داده می‌شود و جدید بودن رویکرد اتخاذ شده با توجه به چهارچوب‌های از قبل موجود، تأکید می‌شود. خلاصه‌سازی انتزاعی با استفاده از گراف‌ها و رویکردهای مختلف یادگیری ماشینی در گذشته پیشنهاد شده است.

خلاصه‌های ایجاد شده با استفاده از نمایش معنایی متن در قالب نمودار به‌عنوان نمودارهای نمایشی به معنای انتزاعی (AMR : Abstrac Meaning Representation) نامیده می‌شود.

1. Bhargava et al.

جملات ورودی برای ایجاد نمودارهای AMR، جداگانه تجزیه می‌شود؛ نمودارهای جداگانه با استفاده از الگوریتم پیش‌بینی مدل پرسپت رون به نمودار خلاصه تبدیل می‌شود .

سپس با پیش‌بینی و انتخاب زیر گراف‌ها با دقت بالا بارگاوا و همکارانش تولید می‌شود آن‌ها رویکردی را پیشنهاد کردند که از نمودارهای جهت‌دار استفاده می‌کند و از ترتیب کلمات جمله اصلی برای ایجاد مفاهیم انتزاعی استفاده می‌کند .

رویکرد پیشنهادی با استفاده از فرم گرافیکی متن ورودی، افزونگی را کاهش می دهد. اگر دو جمله جمع‌شونده باشند، از رابط‌ها براساس احساسات جملات مجاور استفاده می‌شود. در گذشته نه چندان دور ، مدل‌های توالی به دنباله در بسیاری از مشکلات مانند ماشین موفق بودند (Bhargava et al., 2020).

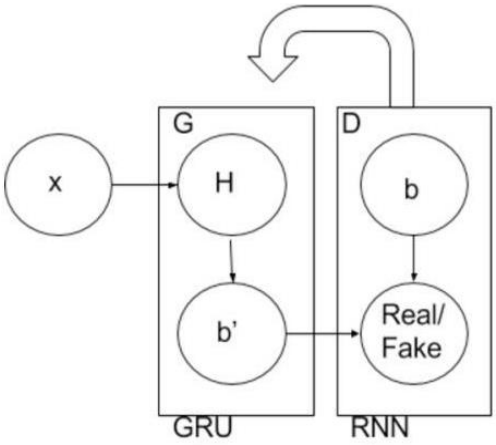
فخرزی و همکاران[1] (۲۰۲۱)، در این مقاله به بررسی کتاب تفسیرالمصباح آثار م.قریش شهاب پرداخته‌ایم؛ این کتاب، تفسیری از زبان اندونزیایی استفاده می‌کند که با سبک نوشتاری ارائه شده است که به راحتی توسط همه افراد از دانشگاه گرفته تا جامعه قابل هضم است و معنای یک آیه نوشته شده را آن‌طور توضیح می‌دهد که خواننده را برای مطالعه جذب می‌کند. منابع تفسیر مصباح، قرآن، حدیث، تفسیر اجتهادی شهاب قریش و نقل از مفسران دیگر است. در این مقاله درباره‌ی بردارهای توزیعی که کلمات را نشان می‌دهند بحث می‌شود. تبدیل عبارات به بردار برای کاربردهای پردازش زبان طبیعی، مفید است؛ این فرایند با فعال کردن محاسبه‌ی کلمات مرتبط معنایی انجام می‌شود که سپس برای نشان دادن سایر واحدهای زبانی استفاده می‌شود. مطالعه بردارهای کم‌بُعد متون، اجسام به‌عنوان جایگزینی با استفاده از شبکه‌های عصبی قابل دستیابی است، این روش برای دستیابی به عملکرد خوب در NLP تعیین می‌شود (Fakhrezi et al., 2021).

ساهو و همکاران[2] (۲۰۱۸)، این بخش بیشتر در مورد خلاصه‌سازی انتزاعی بحث می‌کند. SUMMONS، یک سیستم خلاصه مفهومی و زبانی است که مجموعه‌ای از الگوها خلاصه می‌کند که حاوی ویژگی‌های برجسته موجود در مقاله ورودی است، اما سیستم، خلاصه‌هایی تولید می‌کند که فقط از یک تا سه جمله تشکیل شده‌اند. رجینا بارزیلای و همکارانش یک مدل مبتنی‌بر نمودار، پیشنهاد کرده‌اند که اشتراک و تنوع را در بین جملات خوشه‌ای با استفاده از هم‌ترازی چنددنباله‌ای (MSA :Multiple Sequence (Alignment پیدا می‌کند هنگام محاسبه‌ی امتیاز MSA بین یک جفت حمله فقط اشتراکات سطح کلمه در نظر گرفته می‌شود، اما رابطه دستوری بین جملات برای محاسبه‌ی امتیاز MSA در نظر گرفته نشده است.

بانک اطلاعاتی

این مجموعه داده، شامل صد مقاله خبری از انگلیس، هند، مالزی و غیره در میان ۱۰۰ سند، ۱۰ مقاله خبری مربوط به یک موضوع است؛ ۱۰ سند در یک سند ترکیب می‌شود تا به‌عنوان ورودی به رویکرد پیشنهادی در این کار ارائه شود.

1. Fakhrezi et al.
2. Sahoo et al.

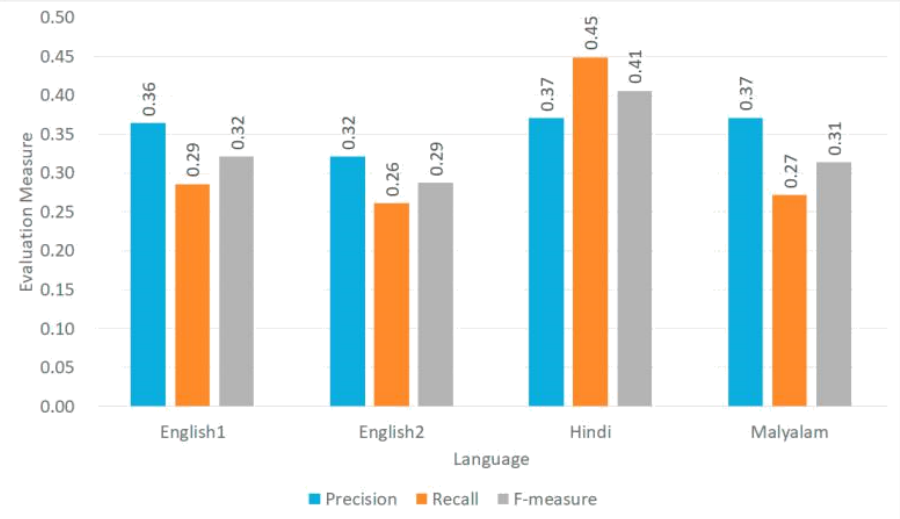


Algorithm 1 Text Summarization using paraphrase detection

```
Input: A document D to be summarized.
Output: An output Summ, the summary.
Initialisation: Sen=[]
Preprocessing
Set Sen, which contains the sentences already processed.
for Sentence S in D do
    flag = true
    for S' in Sen do
        if S' and S are paraphrases then
            flag = true
        end if
    end for
    if flag == false then
        Add S to Sen
    end if
end for
Sen is provided as the input to GRU
Train the GAN, generate Probability distribution P
sentence = ""
while Summ <required length  do
    while sentence does not specify grammar rules and len(sentence) <10 do sentence.append(predict(sentence))
    end while
    if sentence satisfies grammar rules then
        Add sentence to Summ
    end if
end while
```

این مجموعه شامل داده بررسی‌های محصول به زبان انگلیسی از سیستم GPS ناوبری، اتومبیل‌ها و ipod است؛ ۵۱ سند مربوط به این نظرات کاربران وجود دارد و هر بررسی به‌طور متوسط ۱۰۰ جمله دارد.

در مدل RNN پیشنهادی، تعداد لایه‌ها به دو عدد ثابت می‌شود؛ اندازه دسته برای RNN و GRU ۵۰ است. نرخ پوسیدگی ۹.۰ است و نرخ یادگیری از ۰۰۲.۰ شروع می‌شود، بهینه‌ساز ADAM برای تعیین نرخ یادگیری تطبیقی استفاده می‌شود. آموزش برای ژنراتور زمانی‌که تلفات کمتر از ۵.۰ باشد و برای تفکیک‌کننده زمانی‌که تلفات کمتر از ۳.۰ باشد متوقف می‌شود (Bhargava et al., 2020).



و اما مجموعه داده‌های مورد استفاده در این پژوهش شامل پرس‌وجو، تفسیرالمصباح و خلاصه‌سازی هدف بود. String Query ورودی سیستم به‌صورت کلماتی است که می‌خواهند معنی پیدا کنند که یک واژگان است و دارای ۵۰ کلمه است که برای معنی جست‌وجو می‌شود.

کتاب تفسیرالمصباح به سطح پاراگراف تقسیم شده و سپس به txt دیجیتال شده و دارای ۵۰۰ سند برای استفاده به‌عنوان معانی پرس‌وجو است و به هر سند یک شناسه براساس شماره سوره‌ها و آیات داده می‌شود. به‌طور مثال جمله اول سوره انعام آیه ۱۸ آدرس ۱.txt–۱۸–۶ را می‌گیرد.

خلاصه‌سازی مورد انتظار به کمک سیستم و به‌صورت دستی پردازش می‌شود که معمولاً از ۵۰ سند باز استخراج می‌شود، پس از وارد کردن مجموعه داده‌ها، پیش پردازش را با استفاده از نشانه‌گذاری‌ها انجام می‌دهیم، کلمات اضافی را حذف و متد را در پیش می‌گیریم:

Before Preprocessing	After Preprocessing
Tatkala perjalanan matahari telah melampaui pertengahan dan telah menuju kepada terbenamnya dinamai ashr / asar.	'tatkala', 'jalan', 'matahari', 'lampau', 'tengah', 'benam', 'nama', 'ashr', 'asar'

برای محاسبات شباهت کسینوس، می‌توان با تبدیل جمله به بردار، آن را محاسبه کرد زیرا در این صورت شباهت کسینوسی قابل محاسبه است.

Before Vectorization	After Vectorization
'tatkala', 'jalan', 'matahari', 'lampau', 'tengah', 'benam', 'nama', 'ashr', 'asar'	'tatkala': 1, 'jalan': 1, 'matahari': 1, 'lampau': 1, 'tengah': 1, 'benam': 1, 'nama': 1, 'ashr': 1, 'asar': 1

سپس نتایج به‌دست‌آمده را با تمام اسناد مقایسه کنید تا شباهت‌های بین نتایج و اسناد را ببینید و نتایجی که بیشترین شباهت به اسناد را دارند مرتب کرده که معمولاً بین ۵۰۰ سند، حداقل ۱۰ و حداکثر ۱۵ سند مشابه است (Fakhrezi et al., 2021).

در شکل زیر یک نمونه محاسبه شباهت کسینوسی را مشاهده می‌کنید:

Documents	Score Cosine Similarity
'67-2-10.txt'	0.7132
'67-2-7.txt'	0.6969
'67-2-5.txt'	0.4472
'67-2-6.txt'	0.4472
'67-2-9.txt'	0.3481
'67-2-4.txt'	0.3333
'67-2-3.txt'	0.3015
'2-132-2.txt'	0.2887
'67-2-8.txt'	0.2673
'75-34-9.txt'	0.2425
'75-34-10.txt'	0.1643

رویکرد ترکیبی به روش خلاصه‌سازی چکیده، یک سند متنی را به‌عنوان ورودی خود می‌گیرد و خلاصه فشرده و انتزاعی متن را تولید می‌کند (Sahoo et al., 2018).

این رویکرد دارای ۵ مرحله است:

۱- پیش‌پردازش و خوشه‌بندی جملات

۲- رتبه‌بندی جمله

۳- ادغام جمله مبتنی‌بر قانون

۴- فشرده‌سازی جمله

۵- ارائه خلاصه نهایی

سند ورودی به صورت یک گراف بدون جهت G=(V,E) نمایش داده می‌شود؛ جایی که V در مجموعه نشان دهنده جملات و E مجموعه یال‌هایی است که قدرت بین یک جفت جمله را نشان می‌دهد که یک ویژگی مهم برای خوشه‌بندی جملات است در مقایسه با یک سند، یک جمله تأثیر کمتری دارد

بنابراین رابطه سه جمله‌ای برای خوشه‌بندی جمله در نظر گرفته می‌شود.

روابط سطح جمله عبارت اند از:

۱- رابطه انتقال

۲- رابطه آنافوریک

۳- اصطلاحات مشترک بین یک جفت جمله

وزن رابط گذار

اگر رابطه انتقال صریح از بین دو جمله sj و si خارج شود، یک تقویت‌کننده انتقال ضریب TB به لبه E دو جمله داده می‌شود. اگر فرض کنیم که برای هر جمله ۲ - i<=N رابطه انتقال صریح ممکن است با دو جمله بعدی وجود داشته باشد، بنابراین یک ماتریس وزن انتقال با بعد N*N ایجاد می‌شود.

وزن رابط آنافوریک

رابطه آنافوریک، نقش مهمی مانند رابطه گذار در خوشه‌بندی جمله ایفا می‌کند؛ همچنین فرض می‌کنیم که برای یک جمله رابطه آنافوریکی ممکن است با دو جمله بعدی آن وجود داشته باشد؛ ضریب تقویت آنافوریک AB به یک یال داده می‌شود ماتریس وزنی آنافوریک نیز با ابعاد N*N ایجاد می شود.

وزن رابطه مشترک

تعداد اصطلاحات متمایز مشترک بین دو جمله si وsj به‌عنوان قدرت تشابه نام‌گذاری شده است. یک ضریب تقویتی B ضرب در تعداد مشترک عبارت k بین یک جفت جمله اختصاص داده می شود. TSWij یا وزن مدت مشابه (term similarity weight) بیانگر قدرت تشابه بین دو جمله است که در ابعاد N*N ارائه می‌شود (Sahoo et al., 2018).

نتیجه‌گیری

در آخر، رویکردی در این مقاله برای خلاصه‌سازی انتزاعی متن با استفاده از یک شبکه متخاصم مولد برای انجام خلاصه‌سازی متن چندزبانه پیشنهاد شده است. بهبودهایی را می‌توان از نظر تنظیم فراپارامترها و اندازه مجموعه داده انجام داد؛ برای نتایج بهتر، ورودی مورد استفاده به شبکه آموزش بهتر است بزرگ‌تر باشد. زیرا دستیابی به مجموعه داده‌های بزرگ، برای بهبود مدل یکی از کارهای اساسی است. تکنیک بازنویسی مورد استفاده هنوز یک زمینه تحقیقاتی فعال در زبا ن هندی است و از آنجایی که این تکنیک برای حذف افزونگی به همان روش تکیه می کند، می‌توان آن را برای ارائه نتایج مطمئن‌تر بهبود بخشید (Bhargava et al., 2020). همچنین می‌توان با کمک تکنیک TextRank متن یا اسناد را خلاصه کرد؛ روش کار این تکنیک به این صورت است که تمام کلمات را معنی، سپس خلاصه‌ای براساس تمامی معانی کلمات به دست آمده ارائه می‌دهد، در مرحله بعد با مقایسه خلاصه‌های به دست آمده با اسناد یا جملات اصلی، مشابه‌ترین خلاصه را انتخاب کرده و نتیجه خلاصه‌سازی سیستم را با خلاصه‌سازی دستی(توسط انسان) مقایسه می کند. نتایج خلاصه‌سازی خودکار متن با استفاده از روش TextRank تکراری نیست و برای برخی از پردازش‌ها تقریباً با خلاصه‌ای که به صورت دستی توسط انسان ایجاد شده یکسان است (Fakhrezi et al., 2021). همچنین برای خلاصه‌سازی انتزاعی متن می‌توان خوشه‌بندی جملات مرتبط براساس یک اصل زبانی انجام داد، بدین معنی که رابط تشبیهی، رابطه انتقالی، جایگاه جمله و داشتن اسم خاص در بین جمله‌ها انجام می‌شود. پس از خوشه‌بندی، جملات با رویکرد آماری رتبه‌بندی می‌شوند؛ بهترین جمله برای ادغام با جملات برتر از هر خوشه پیدا می‌شوند. دقت مدل سیستم فشرده‌سازی در این روش حدود ۷۱ درصد است، بنابراین زمینه بیشتری برای بهبود دقت سیستم فشرده‌سازی جملات وجود دارد، به‌طوری که عملکرد کلی سیستم خلاصه‌سازی بهبود می‌یابد (Sahoo et al., 2018).

منابع

Bhargava, R., Sharma, G., & Sharma, Y. (2020). Deep text summarization using generative adversarial networks in Indian languages. *Procedia Computer Science*, *167*, 147-153.

Fakhrezi, M. F., Bijaksana, M. A., & Huda, A. F. (2021). Implementation of automatic text summarization with TextRank method in the development of Al-qur'an vocabulary encyclopedia. *Procedia Computer Science*, *179*, 391-398.

Sahoo, D., Bhoi, A., & Balabantaray, R. C. (2018). Hybrid approach to abstractive summarization. *Procedia computer science*, *132*, 1228-1237.