همایش ملی پیشرفت های تکنولوژی در مهندسی برق، الکترونیک و کامپیوتر
National E-Conference of Technology Developments on Electrical, Electronics and Computer Engineering
. . . W W W . T D E C O N F . I R . . .

بررسی روش های خلاصه سازی خودکار متن

سمیرا بابااحمدی میلانی ، نسیم نور افزا
دانشگاه آزاد اسلامی واحد زنجان
دانشگاه آزاد اسلامی واحد نجف آباد
مسئول مکاتبات : سمیرا بابااحمدی میلانی

چکیده:
فرایند کوتاه کردن یک منبع به صورتی که حاصل حاوی اطلاعات مهم آن باشد را خلاصه سازی می گویند. .برای بشر خلاصه سازی مدارک واطلاعات به صورت دستی کار مشکلی است.خلاصه سازی متون منجر به استفاده از منابع بیشتر با سرعت بالاتر و در نتیجه حاصل شدن اطلاعات غنی تر میشود .روش های خلاصه سازی متن به دو گروه استخراجی وچکیده ای تقسیم می شوند.روش خلاصه سازی استخراجی، شامل مهم ترین جملات وپاراگراف های مدارک ومطالب می باشد و اهمیت یک جمله بر اساس روش های آماری و ویژگی زبان شناسی آن مشخص می شود.روش خلاصه سازی چکیده ای شامل مفهوم متن اصلی است و بازگویی کلمات در آن کمتر می باشد که از روش های زبانشناسی برای آزمودن و تفسیرمتن سپس برای پیدا کردن مفاهیم جدید استفاده می شود و همچنین به پیدا کردن مفاهیم جدید و پیدایش یک متن کوتاه تر کمک می کند واغلب مطالب مهم متن اصلی را در بر دارد.اکثر سیستم های خلاصه سازی کنونی از نوع استخراجی می باشند.

کلمات کلیدی : خلاصه سازی خودکار متن ، چکیده سازی ، خلاصه سازی استخراجی جملات ،روش های خلاصه سازی متن

۱-مقدمه

خلاصه سازی خودکار متن [Jezek,Steinberger,2008]دارای اهمیت ویژه ای است و یک ابزار زمانی برای دست یابی وتفسیر اطلاعات یک متن است که در دهه های اخیر پیشرفت زیادی داشته است . برای بشر خلاصه کردن اسناد واطلاعات زیاد به صورت دستی کاربسیار مشکلی است. مقدار زیادی از مواد متنی در اینترنت قابل دسترس است. اگر چه معمولا اینترنت اطلاعاتی بیشتر از آن چیزی که می خواهید به شما می دهد بنابراین شما با مشکل بیشتری مواجه می شوید .جستجو برای اسناد مربوطه در میان تعداد زیادی از مطالب قابل دسترس و مربوط به موضوعی که ما می خواهیم کار دشواری است هدف خلاصه سازی خودکار متن، فشرده کردن منابع متنی به نسخه های کوتاهتر که شامل اطلاعات مهم متن و یک مفهوم کلی نیز می باشد.یک خلاصه سازی متون منجر به استفاده از منابع بیشتر با سرعت بالاتر و در نتیجه حاصل شدن اطلاعات غنی تر میشود [کریمی ، شمس فرد، ۱۳۸۵] یک خلاصه [Fan,Wallace,2005] را می توان به عنوان یک راه اشاره ای به برخی از قسمت های اسناد و اطلاعات اصلی استفاده کرد.یا یک راه خبری برای پوشش تمام اطلاعات مربوط به متن خبری می باشد.مزیت اصلی خلاصه سازی ،کاهش زمان مطالعه است. یک سیستم مناسب خلاصه کردن باید عنوانهای مختلف یک متن را منعکس کند و مقدار حشو ناچیزی را داشته باشد. همچنین خلاصه سازی ابزاری برای تحقیق درباره ی موضوعات مختلف است و نشانه هایی در موضوعات فرعی به منظور مشخص کردن نکات کلیدی اطلاعات و اسناد می باشد. نرم افزار خودکار خلاصه سازی کلمات مایکروسافت یک نمونه ساده ای از خلاصه سازی متن می باشد. روش خلاصه سازی متن به دو گروه خلاصه سازی استخراجی و خلاصه سازی چکیده ای تقسیم می شود.روش خلاصه سازی استخراجی [Erkan,radev,2004] [Han,romacker,2001] شامل انتخاب جملات و پاراگراف های مهم از اسناد و متون اصلی می باشد. و آنها را کوچکتر می کند. اهمیت جملات بر اساس آمار و ویژگی زبانشناسی آنها مشخص می شود.در یک خلاصه سازی چکیده ای [Kyoomarsi,khosravi,2008] تلاش می شود درک مفهوم اصلی اسناد و اطلاعات حفظ شود و همچنین این مفاهیم به زبان ساده بیان شوند.که در آن از روش های زبانشناسی برای آزمودن و تفسیرمتن سپس برای پیدا کردن مفاهیم جدید استفاده می شود . و برای بیان بهترین توصیف ارائه یک متن کوتاهتر که اغلب مطالب مهم متن اصلی را دربر داشته باشد ضروری است. در فرایند خلاصه سازی چکیده ای به استخراج ساده جملات اکتفا نکرده و در واقع مفهومی از متن اصلی را خواهیم داشت که بسیار نزدیک به مدل ذهنی انسان می باشد. در تولید چکیده پاره -ای از جمله ها یا همه ی آن ها بازنویسی میشوند به عنوان مثال ، اگر در یک خلاصه ،جمله ی (او سیب و انگور و گیلاس ها را خورد) به صورت(او میوه ها را خورد) نوشته شود آن خلاصه یک چکیده است. فرایند خلاصه سازی چکیده ای فرایندی بسیار پیچیده و دشوار است چرا که نیازمند نمایشی مفهومی از متن می باشد و رسیدن به این نمایش بسیار مشکل خواهد بود. علاوه بر این ساخت جمله ای جدید نیاز به اطلاعات زبان شناسی بسیار قوی می باشد[مشکی،آنالویی،۱۳۸۸] .در خلاصه سازی استخراجی ، از متون اولیه چندین بخش از متن را انتخاب

کرده و بر اساس یک معیار اولویت آن ها را مرتب می نمایند. در این نوع خلاصه سازی هر جمله از متن یا متون اولیه رونوشت می شود و جملات متن تغییرات نحوی و معنایی ندارند .

خلاصه سازی استخراجی بدین گونه شکل می گیرد [کیومرثی ،۱۳۸۷] :بخش های کلیدی استخراج شده از متن (جمله ها و صفحات..) بر اساس تحلیل آماری از ویژگی های سطحی تکی یا ترکیبی مانند تکرار عبارت ها و کلمات, موقعیت کلمات کلیدی ،طول جملات ، اسامی خاص و استخراج شده اند .این گونه رویکرد ها نتیجه اجتناب از درک عمق متن می باشد.که اینها مفاهیم ساده ای می باشند و به راحتی قابل اجرا هستند.روند خلاصه سازی متن [Gupa,Lehal,2009] را می توان به دو مرحله تقسیم کرد: ۱-مرحله پیش پردازشی ۲-مرحله پردازش .

پیش پردازش شامل:

الف-شناسایی مرز جملات و واژه ها:در زبان انگلیسی معمولا مرز جمله با قرار گرفتن نقطه در پایان جمله مشخص می شود.در زبان فارسی به دلیل مشکلات خاص زبان،مانند تشخیص افعال مرکب و کسره ی اضافه این کار به آسانی زبان انگلیسی صورت نمی گیرد.

ب-حذف کلمات غیر مهم: کلماتی که معمولا معنای خاصی ندارندو اگر حذف شوند در معنای اصلی خللی ایجاد نمی کند مانند حرف ربط ، حرف اضافه ،افعال ربطی واکثر افعال ساده و قیود. از متن حذف می شوند تا برای امتیازدهی به جمله به حساب نیایند.

ج- ریشه یابی: هدف ریشه یابی مشخص کردن ریشه و پایه هر کلمه می باشد که بر معنای آن تاکید دارد.در مرحله پردازش ویژگی تاثیر گذار بر ارتباط جملات انتخاب و حساب می شود و سپس بار آموزشی آنها را در نظر خواهیم گرفت . در نهایت از ویژگی بار جملات برای امتیاز دهی به آنها استفاده می شود.وجملاتی که در رتبه بالاتری قرار بگیرند برای خلاصه نهایی انتخاب می شوند.مشکلاتی که در رابطه با خلاصه سازی استخراجی وجود دارد به شرح زیر می باشد.[Lin, 2009][Cheung,2008]

۱-جملات استخراجی معمولا بلندتر از حد معمول اند، قسمتی از این بخش ها برای خلاصه ضروری نیستند , همچنین وقفه های زیادی را بوجود می آورند .

۲-اطلاعات مربوطه و مهم معمولادر بین جملات پخش می شوند وخلاصه سازی استخراجی نمی تواند آنها را جمع کند(اگرچه خلاصه به اندازه کافی طولانی است که بتواند این جملات را در خود جای دهد.)

۳-ناسازگاری اطلاعات مانع از ارائه درست می شود.

۴-یک خلاصه سازی استخراجی خالص اغلب به مشکلاتی در رابطه با ارتباط کلی خلاصه منجر می شود-مسئله مهمی که وجود دارد ،تکرار یک یا چند عبارت متوالی« افزونگی» می باشد. جملات اغلب شامل ضمایر هستند زمانی که از متن استخراج می شوند مرجع خود را از دست می دهند. بدتر از آن, پیوند دادن مفاهیم نامربوط به یکدیگر ممکن است به تفسیر نادرست منجر می شود.(نتیجه ارائه نادرست از منبع اطلاعات) مسئله مشابهی نیز در مورد بیانات موقتی وجود دارد.این مسئله ها در مواردی که چند نمونه از اسناد وجود دارند بیشتر است.بنابراین استخراج از منابع مختلفی گرفته می شود.رویکرد کلی که به این مسائل می پردازد استخراج پس- پردازش می باشد. برای مثال جایگزینی ضمیر با مرجع ضمیر و جایگزینی کلمات زمانی مرتبط با بیان زمان اصلی و غیره.

مسائل و مشکلاتی که در ارتباط با خلاصه سازی چکیده ای وجود دارد به شرح زیر می باشد:[Lin,2009]

بزرگترین چالش درمورد خلاصه سازی چکیده ای مشکل ارائه آن می باشد.سیستم هایی وجود دارند که می توانند آن هارا مجبور کنندکه ارائه قویتری داشته باشندو همچنین در کلی سازی این ساختارها توانایی دارند.(وقتی ارائه ای درست نباشد سیستم ها قادر به خلاصه سازی نیستند .)در مقیاس کمتر،این امکان برای برخی دستگاهای ساختاری مناسب نیز وجود دارد. اما هدف کلی و راه حل ،به تحلیل معنایی در مقیاس وسیعی بستگی دارد .سیستمی که بتواند به درستی زبان طبیعی را بفهمد، از توانایی تکنولوژی امروزی خارج است .ارزیابی خلاصه [Nenkova,Passonneau,2004][Lin,2004][Hovy,Lin,2006],یک جنبه مهم از خلاصه سازی متن می باشد. به طور کلی خلاصه با دو مقیاس اصلی و فرعی ارزیابی می شود .روش های اصلی با مقایس های خلاصه ای کیفی با استفاده از ارزش یابی بشری به کار گرفته می شود. وروش های فرعی مقیاس مشابه ای دارند که برپایه فعالیت و عملکرد اندازه گیری می شود. Newsblaster یک مثال مناسب برای خلاصه سازی متن می باشد که به کاربران کمک می کند اخباری که علاقه دارند را پیدا کنند .این سیستم به طور خودکار اخبار را جمع آوری و سازمان دهی می کند .اخبار را روزانه از سایت های مختلف (CNN,FOX NEWS,ECT) اینترنت گرفته وخلاصه سازی می کند و به کاربران فرصت یک استفاده دوستانه و مناسب را می دهد .

۲-تاریخچه اولین خلاصه سازی متون

علاقه مندی به خلاصه سازی خودکار از اوایل قرن پنجاه میلادی بوده است. یکی از مهمترین صفحاتی که در آن روزها خلاصه شده است :در سال ۱۹۵۸میلادی پیشنهاد شده بود که وزن جملات یک سند را بر اساس تکرار کلمات آن در نظر بگیریم[Luhn,1958].در سیستم خلاصه سازی خودکار متن [Edmondson,1969] علاوه بر روش استاندارد کلمات کلیدی , از سه روش دیگر نیز استفاده می شد.

۱- روش اشاره ای یا راهنمایی: این روش بر اساس فرضیه ایست که ارتباط جملات و اهمیت آنها بر اساس کلمات اشاره ای مشخصی که در دایره لغات کلمات وجود دارد مشخص می شود .

۲- روش عنوانی :در این روش وزن جملات بر اساس مجموع تمام کلمات متن که مربوط به عنوان و عنوانهای کوچک متن اند مشخص می شوند.

۳- روش مکانی: این روش بر اساس این فرضیه می باشد که جملاتی که در ابتدای متن یا در پاراگراف های مجزا قرار دارند , دارای اهمیت بیشتری هستند.

نتایج نشان میدهد که بهترین ارتباط بین استخراج خودکار وبشری , با استفاده از ترکیب این سه روش بدست می آید.خلاصه ساز متن Trainable , جملات را بر اساس وزن اکتشافی آنها انتخاب می کردند [kupice,1995] که از ویژگی های زیر استفاده و آنها را ارزیابی میکردند.

۱ -ویژگی طول جملات: جملاتی که تعداد کلماتشان از حد مشخص کمتری بود برای چکیده استفاده نمی شدند.

۲-ویژگی عبارت های ثابت: جملاتی که دارای عبارات و کلمات خاصی بودند شامل خلاصه می شدند.

۳-ویژگی پاراگرافی: این ویژگی اساسا با ویژگی روش مکانی [Edmondson,1969] برابری می کند.

۴-ویژگی کلمات ریشه ای: اغلب کلمات پر تکرار به صورت کلمات ریشه ای تعریف می شوند.وجملات بر اساس کلمات ریشه ای امتیازدهی می شوند.

۵-ویژگی کلمات مهم: ویژگی کلمات مهم تقریبا مانند کلمات ریشه ای است.

مجموعه ای از اسناد که در آنها از این روش استفاده شده است شامل: ۱۸۸ سند خلاصه بر گرفته از ۲۱ مرکز نشر در زمینه های علمی و فناوری بوده است. این خلاصه ها توسط کارشناسان حرفه ای بوجود آمده است و جملات به کار رفته در خلاصه ها برگرفته از متن اسناد اصلی بودند.

سیستم استخراجی متن Anes , [Brandow,Mitza,Lisa,1995] یک سیستم خودکار می باشد. که حجم وسیعی از اطلاعات خبری را دارا می باشد.روند کلی این خلاصه سازی ۴ جزء اصلی دارد که به شرح زیر می باشد:

۱-تحلیل مجموعه ای از نوشتجات : عمدتا شامل محاسبه tf-idf برای مشخص کردن وزن تمام کلمات می باشد.

۲-انتخاب آماری کلمات مشخص : شامل کلمات با وزن tf-idf بالا هستند.

۳- وزن جملات: مجموع کل وزن کلمات تشکیل دهنده با وزن برخی از فاکتور های دیگر مانند موقعیت مربوطه

۴- انتخاب جملات :جملاتی که امتیاز بالایی دارند:[Mittendorf,1994]

خوشه بندی [Bookstein,1995]: ساختن پیوند و زنجیره ها و خوشه بندی اصطلاح ها و کلمه ها و عبارت ها و بقیه قسمت های شاخص اسناد با استفاده از اطلاعات استاندارد به کار گرفته می باشد.اگرچه این موضوع مسئله ی بزرگی در سیستم های خلاصه سازی چکیده ای ایجاد نمی کند.ولی هنگام ساختن چنین سیستم هایی درنظر گرفتن این موضوع اهمیت دارد.

۳-ویژگی هایی برای خلاصه سازی استخراجی متن

برخی ویژگی ها که برای انتخاب یک جمله در خلاصه نهایی بررسی میشوند عبارتند [Kiyomarsi,2008][chn,2002][Rasim,2005] :

الف: ویژگی محتوای کلمات: کلمات محتوایی یا کلمات کلیدی معمولا اسم ها هستند.جملاتی که دارای کلمات کلیدی اند از شانس بیشتری برخوردارند که جزء خلاصه باشند.

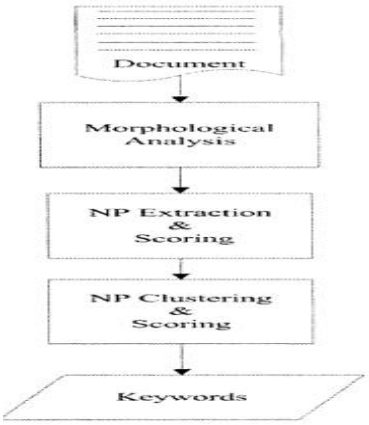
روش های دیگر برای استخراج کلمات کلیدی به شرح زیر می باشد [Gupta ,2009][Bracewell,2005]:

۱-تحلیل لغت شناسی

۲-استخراج عبارت های اسمی وامتیاز دهی

۳-دسته بندی عبارت های اسمی و امتیاز دهی

شکل ۱ روش استخراج لغات کلیدی را به صورت را به صورت شکل ارائه داده است.



شکل ۱- روش استخراج لغات کلیدی

ب- ویژگی کلمات عنوانی : این جملات شامل کلماتی اند که در عنوان ها قرار دارند و مضمون مطلب را مشخص می کنند. جملاتی که حاوی این لغات اند برای خلاصه مناسب می باشند.

پ-ویژگی جایگاه جملات : معمولا اولین و آخرین جمله از اولین و آخرین پاراگراف یک متن اهمیت بیشتری دارند. و برای خلاصه مناسب اند.

ت-ویژگی طول جملات : جملات بیش از حد طولانی و بیش از حد کوتاه در خلاصه استفاده نمی شوند.

ث-ویژگی اسامی خاص: اسم خاص شامل نام یک شخص و مکان و مفاهیم و غیره می باشد. جملاتی که دارای اسامی خاص هستند برای خلاصه مناسب اند.

ج- ویژگی کلمات برتر: جملاتی که شامل سرنام ونام های خاص هستند شامل این گروه می شوند.

چ-ویژگی عبارت های کلیدی: جملاتی که شامل هر عبارت کلیدی اند (برای مثال:در نتیجه, این نامه, این گزارش , خلاصه, بحث ,هدف,گزارش ,تلاش و غیره) برای خلاصه مناسب هستند.

ح-ویژگی کلمات پایه: اگر کلمه ای که در جمله می باشد جزء لیست کلمات پایه و ریشه ای باشد پس این جمله مهم می باشد.لیست کلمات پایه قبلا توضیح داده شده است و می توان شامل تعدادی از کلمات خاص باشد.

خ- ویژگی فونت پایه: جملاتی که شامل کلماتی اند که پررنگ تر اند یا زیر آن خط کشیده شده است یا با حروف خاص مشخص شده اند مهم میباشند.

د- ضمایر: ضمایری مانند (او ،آنها ،آن) را نمی توان در خلاصه استفاده کرد مگر اینکه به جای این ضمایر اسم ها و مرجع آنها جایگزین شود.

ذ- ارتباط جمله به جمله :میزان شباهت جمله S با دیگر جملات سند محاسبه می شود. سپس مقادیر شباهت آن ها اضافه می شود و یک مقدار خام از این ویژگی برای جمله S بدست می آید وبه همین ترتیب این روند برای تمام جملات تکرار می شود.

ر- ارتباط جمله با مرکز ثقل: برای هر جمله S بردار نماینده مرکز ثقل هر سند محاسبه میشود سپس شباهت بین مرکز ثقل و هر جمله حساب می شود و مقدار اولیه این ویژگی برای هر جمله بدست می آید.

ز-وجود اطلاعات غیر ضروری : برخی کلمات اطلاعات غیر ضروری را مشخص می کنند. این کلمات نشانه های سخنی اند مانند :زیرا , علاوه براین و غیره می باشند و معمولا اول جمله می آیند. این جملات ویژگی دوتایی دارند، که اگر یکی از این کلمات را داشته باشند درست و در غیر این صورت غلط هستند.

ژ- تحلیل مباحثه: سطح اطلاعات سخن در متن [Chan,2000], یک ویژگی مناسب برای خلاصه کردن است. به منظور به وجود آوردن یک ارتباط و خلاصه روان وغیره ، ضروری است که ساختار کلی مباحثه را در متن در نظر بگیریم و جملات حاشیه ای را حذف کنید.

این ویژگی ها مانند تعدادی از روش های خلاصه سازی مهم می باشند و همچنین شامل مشخصه های آماری و زبانشناسی هستند.

۴- روش‌های خلاصه‌سازی:

روش های خلاصه سازی خودکار متن از دیدگاههای مختلفی تقسیم بندی می شوند .این دسته بندی ها را از چند جنبه می توان شرح داد:

معیارهای ورودی: شامل اطلاعات آغازین به همراه نمایش گروه زیادی از معیارهای ورودی که به عوامل زیر تقسیم می‌شود:

- زبان متن: بیانگر اینکه روش خلاصه‌سازی بصورت تک زبانه یا چند زبانه باشد متفاوت می‌باشد. به دلیل اینکه زبان‌ها با یکدیگر متفاوتند و ویژگی‌های متفاوتی بر هر کدام حکم فرماست که ممکن است بر دیگری برقرار نباشد.

- منبع: منابع ورودی در روش خلاصه‌سازی می‌تواند تک سندی یا چند سندی باشد.

خلاصه سازی تک سندی چون در خلاصه سازی تک سندی ورودی سیستم یک سند میباشد[Mani,1999] پیچیدگی خلاصه سازی تک سندی به مراتب کمتر ازخلاصه سازی چند سندی می باشد.تاکنون روش های زیادی برای خلاصه سازی تک سندی ارائه شده اند که تعدادی از آن ها در مراجع [SvoRE,2007][Mihalcea,2005] ذکر شده است.

خلاصه سازی چند سندی عبارت است از تولید یک خلاصه از محتوای اطلاعاتی یک مجموعه ی سنددرباره ی یک موضوع اصلی .مهم ترین چالش ها در خلاصه سازی چند سندی عبارتند از [wan,20007]

خوانایی:چون اسناد با دیدگاههای مختلفی به شرح یک موضوع می پردازند گاها متناقض با یکدیگر می باشند.

افزونگی:چون همه اسناد در ارتباط با یک موضوع می باشند بحث همپوشانی جملات انتخاب شده یا همان افزونگی پیش می آید .

ساختار متن: برروی پردازش متن تاثیر مستقیم می‌گذارد و شامل پارگراف، نهاد و گزاره می‌باشد.

هدف خلاصه‌سازی: خلاصه‌سازی به چه صورت انجام پذیرد. با انتخاب موضوع، متن را می‌توان به سه دسته ی آگاهی دهنده برای عموم، اخباری، و خصوصی‌سازی شده توسط درخواست کاربر تقسیم نمود که خود شامل مواردی می باشد از جمله:

- برهان: شرایط یا هدف مشخص می‌کند که خلاصه قرار است در کجا و به چه منظوری باشد. بیان هدف خلاصه‌سازی می‌تواند بصورت هشدار، پیش نمایش، آگاهی، زندگی نامه، موضوعی خاص مربوط به یک سازمان یا انجمن مانند اخبار و گزارش فنی باشد.
- کاربر: خلاصه‌سازی برای همه قابل استفاده است یا مختص به کاربر خاصی است که بستگی به دانش فرد خبره دارد.

- کاربرد: بر اساس اینکه موارد استفاده خلاصه چیست ، نحوه تولید متن خلاصه تغییر خواهد کرد.

نوع خلاصه خروجی : بر این اساس خلاصه‌ها به دو دسته عمده خلاصه‌سازی استخراجی و خلاصه‌سازی چکیده‌ای، تقسیم می شوند[پورمعصومی،۱۳۹۱] [اخوان،شمس فرد،۱۳۸۸]

۵-بررسی روش های خلاصه سازی :

بر اساس یکی از دسته بندی ها [Lin,1998]،روش های خلاصه سازی متن به سه دسته عمده تقسیم بندی می شوند: روش مبتنی بر تکنیک های آماری، روش های مبتنی بر تکنیک های معنایی سطوح بالاتر ،روش های مبتنی بر هوش مصنوعی.

۵-۱ روش های خلاصه سازی آماری

روشهای آماری جزء اولین روشهایی هستند که در خلاصه سازی متن استفاده شده اند. در این روشها عموماً سعی میشود با انجام محاسباتی به جمله وزن داده میشود. سپس جملاتی که بیشترین وزن را دارند به عنوان خلاصه برگردانده میشوند. برخی از این روش ها عبارتند از:

۵-۱-۱ روش فرکانس کلمه

این روش جزء اولین روشهایی است که برای خلاصه سازی استفاده شده است[luhn,1958] و البته هنوز هم به صورت ترکیبی با سایر روش ها استفاده میشود. در این روش در ابتدا با بهره گیری از یکی از روشهای وزن دهی نظیر Tf-Idf به کلمات وزن داده میشود. سپس این وزن در واحد جمله توزیع میشود. در ادامه هم با توجه به میزان فشرده سازی،جملاتی که بیشترین وزن را دارند برای خلاصه انتخاب میشوند. در ادامه به برخی از روشهای وزن دهی به کلمات اشاره شده است. ادعای این روش این است که کلماتی که در متن بسیار پرکاربرد هستند، کلماتی هستند که به موضوع اشار ه می کنند.این روش عموما با حذف ایست واژه ها و ریشه یابی انجام می شود.
در جدول ۱ تعدادی از روش های وزن دهی که مورد استفاده قرار می گیرد آورده شده است:

جدول ۱- تعدادی از روش های وزن دهی

Method	Local weight	Global weight
TF-IDF	$f_{i,j}$	$\log \frac{N}{n_i}$
Log-IDF	$1+\log f_{i,j}$ if $f_{i,j} > 0$ 0 if $f_{i,j} = 0$	$\log \frac{N}{n_i}$
GF-IDF	$\frac{gfi}{ni}$	$\log \frac{N}{n_i}$
No-weight	$f_{i,j}$	None

مشکل این روش این هست که لزوما کلمات پرکاربرد، کلمات مهم نبوده و بسیاری از کلمات هستند که علیرغم کاربرد زیادشان بار معنایی زیادی در متن و در ارتباط با موضوع ندارند و برعکس ممکن است کلماتی باشند که فرکانس رخدادشان کم بوده اما دارای اهمیت بسیار بالایی باشند. لازم بذکر است علی رغم این مشکل، این دسته از روشها در بسیاری از سیستمها به صورت ترکیبی یا ارتقا یافته مورد استفاده قرار می گیرند.

۵-۲-۲ روش مبتنی بر موقعیت یا جایگاه

ایده اصلی این روش این است که جملات مهم متن همواره در جاهای خاص و مشخصی از متن قرار میگیرند. اینکه کدام قسمت از متن دارای اهمیت بیشتری میباشد بستگی به نوع متن دارد. به عنوان مثال برای متون خبری انگلیسی ثابت شده است که در اکثر موارد اولین پاراگراف مهمترین پاراگراف و اولین جمله مهمترین جمله میباشد. یا برای متون علمی و مقالات، مشخصا قسمت چکیده و نتیجه گیری همواره مهمترین قسمت مقاله میباشد.بر همین اساس به جملات بر اساس موقعیتشان در پاراگراف وزن می گیرند. پاراگراف ها هم براساس موقعیتشان در متن وزن میگیرند و نهایتا جملات براساس امتیازشان به عنوان خلاصه انتخاب می شوند. سیستم swesum, [Dalianis,2000]که یک سیستم چندزبانه برای خلاصه سازی اخبار میباشد، بر همین اساس عمل میکند. لازم به ذکراست این قانون برای تمامی زبانها درست نمی باشدیکی دیگر از کاستی های این روش این است که تنها برای نوع خاصی از متون درست عمل می کند و البته در مواردی هم با تغییر ساختار نوشتار و سبک نویسنده کاملا دچار مشکل می شود.

۵-۳-۳روش مبتنی بر عنوان

این روش هم جزء اولین روشهای خلاصه سازی متن میباشد [Edmonson,1969], و ایده اصلی آن این است که همواره موضوع و عنوان متن، بیانگر محتوای آن میباشد و بر اساس همین اصل، جملات مهم با توجه به عنوان متن باید انتخاب شوند. با توجه به این اصل، در این روشها سعی میشود از کلمات موضوع متن به عنوان کلمات کلیدی استفاده شود و از آنها برای استخراج جملات مهم استفاده شود. مشکل اصلی این روش این است که بسیاری از متون فاقد عنوان بوده و یا اینکه امکان استخراج عنوانشان به صورت خودکار فراهم نمیباشد. همچنین ممکن است عنوانشان بسیار کلی باشد مثل عنوان سرفصل های یک کتاب که به صورت کلی به محتوای آن اشاره میکند و یا اینکه تنها تعدادی کلمه اصلی و مهم در موضوع وجود داشته باشد و سایر کلمات غیر اصلی و تنها توصیفی باشند . با این وجود لازم به ذکر است که این روش ساده و بدیهی در بسیاری از روشهای امروزی برای خلاصه سازی متن به صورت ترکیبی استفاده می شود . دربسیاری از روشهای جدید سعی میشود تا در ابتدا موضوعات یا زیر موضوعات پنهان موجود در متن به صورت خودکار استخراج شده و سپس بر اساس آنها جملات مهم انتخاب شوند. از جمله این روشها می توان به روشهای مبتنی بر مدل موضوع اشاره نمود..

۵-۳-۴ روش مبتنی بر عبارات اشاره

بحث اصلی در این روش، عبارات خاصی هست که در تمامی زبان ها وجود دارند و بر اهمیت جمله های بعدشان می افزایند. به این عبارات،عبارت اشاره میگویند. از جمله این کلمات یا عبارات برای زبان انگلیسی به عبارت "the main aim "میتوان اشاره کرد. این عبارات در تمامی زبانها موجود میباشند. در زبان فارسی هم عباراتی نظیر "در این گزارش " ، "به عنوان نتیجه"، "هدف از" ، "به عنوان خلاصه"و"در پایان" جزء عبارات اشاره می باشند. در این دسته از روش های خلاصه سازی، در ابتدا باید کلیه این عبارات و کلمات استخراج شوند. سپس باید گرامری که این کلمات و عبارات در آن صدق میکنند استخراج شوند. آنگاه برای خلاصه سازی ، متن ورودی تجزیه شده و سپس بررسی میشود که آیا در گرامر از پیش تعیین شده صدق میکند یا خیر. جملاتی که دارای عبارات اشاره بوده و در گرامر صدق کنند وزن بیشتری به خود اختصاص خو اهند داد. جملاتی که بیشترین وزن را کسب کنند نهایتا به عنوان جملات برتر برای خلاصه انتخاب میشوند. این روش [Leman,1997]را برای زبان فرانسه پیاده سازی کرد. وی در ابتدا دیکشنری cue,weight را تشکیل داد و سپس گرامر

این زبان را برای زبان فرانسه استخراج کرد و به جملات بر اساس دیکشنری از پیش تعریف شده وزن اختصاص داد. مشکل عمده این روش این است که به تنهایی برای خلاصه سازی کافی نمی باشد چراکه در بسیاری از متون ممکن است عبارات اشاره وجود نداشته باشد یا بسیار کم باشد که در این صورت با مشکل مواجه می شود. از این روش در بسیاری از سیستمهای خلاصه‌سازی امروزی به صورت ترکیبی با سایر روشها استفاده میشود.

۵-۳-۵ روش ترکیبی

ترکیبهای مختلف روشهای ذکر شده هم بررسی شده است. به این ترتیب که به هرکدام از روش های بالا وزنی اختصاص داده میشود. این وزن هم معمولا با استفاده از یادگیری ماشینی بدست می آید. آقای ادمونسون این روشها را با هم ترکیب کرده اند و بهترین حالتی که گزارش دادند استفاده از ترکیب عبارات اشاره بعلاوه عنوان و جایگاه می‌باشد[Edmonson,1969]

۵-۳-۶ روش مبتنی بر دسته بندی کننده های بیزین

در این روشها از قانون احتمال شرطی اولیه بیزین برای خلاصه سازی استفاده میشود [kupice,1995]. بر همین اساس احتمال اینکه جمله ی s در خلاصه ی S قراربگیرد با استفاده از فرمول (۱) بدست می آید

$$P(s \in S \mid F_1, F_2, \ldots F_k) = \frac{P(F_1, F_2, \ldots F_k \mid s \in S) P(s \in S)}{P(F_1, F_2 \ldots F_k)} \quad (۱)$$

که در آن $F_1, F_2, \ldots, F_k$ ویژگی های ۱تا K بوده و می تواند به عنوان مثال هر کدام از روش هلی قبلی باشند. بر اساس قانون بیزین احتمال اینکه جمله ی s در خلاصه ی S قراربگیرد با فرض اینکه ویژگی های $F_1, F_2, \ldots, F_k$ برقرار باشد به صورت رابطه (۲) محاسبه می گردد:

$$P(s \in S \mid F_1, F_2, \ldots F_k) = \frac{\prod_{j=1}^{k} P(F_j \mid s \in S) P(s \in S)}{\prod_{j=1}^{k} P(F_j)} \quad (۲)$$

که همان قانون احتمال شرطی بیزین میباشد. پس از محاسبه این احتمال برای تمامی جملات، جمله هایی که بیشترین احتمال را داشته باشند در خلاصه قرار میگیرند.

۵-۳-۷ روش مبتنی بر زنجیره های مارکوف

زنجیره مارکوف، دنباله ای از متغیرهای تصادفی است که همگی این متغیرهای تصادفی دارای فضای نمونه ای یکسان هستند اما، توزیع احتمالات آنها میتواند متفاوت باشد و در ضمن هر متغیر تصادفی در یک زنجیره مارکوف تنها به متغیر قبل از خود وابسته است. از زنجیره های مارکوف در علوم مختلفی استفاده شده است. در مدیریت و برنامه ریزی و بطور کلی مسائل تصمیم گیری، در شبکه های کامپیوتری و همچنین در بحث خلاصه سازی متن از زنجیره مارکوف استفاده شده است. اولین بار کونری [CONROY,2001] از زنجیره های مارکوف در خلاصه سازی متن استفاده کرده است. وی با استفاده از یکسری ویژگی های داده شده، احتمال حضور هر جمله در خلاصه را محاسبه کرده است. در این مدل نست به روش بیزین، فرض کمتری برای استقلال متغیرهای تصادفی شده است . به عبارت بهتر در مدل مارکوف برای خلاصه سازی، فرض نمیشود که احتمال حضور جمله i ام در خلاصه نهایی مستقل از حضور جمله i-1 در خلاصه باشد (برخلاف روش بیزین). ویژگیهایی که در روش کونری برای توسعه خلاصه‌سازی مبتنی بر زنجیره ای مارکوف ارائه شده است شامل این موارد میباشد: موقعیت جملات در سند، تعداد کلمات موجود در جملات، میزان شباهت کلمات جمله. بر همین اساس در این مدل انتظار می رود که احتمال حضور جمله بعدی در خلاصه بسته به اینکه جمله فعلی در خلاصه حضور داشته باشد یا خیر متفاوت خواهد بود.

مشکلات روش های آماری

تقریبا اکثر روشهای آماری ذکر شده دارای مشکلات مشترکی میباشند. یکی از این مشکلات عدم توجه به کلمات هم خانواده میباشد. به عنوان مثال اگر در متن کلمات ماشین و اتومبیل وجود داشته باشد در روشهای آماری چون مبتنی بر نحو هستند این کلمات جدا از هم در نظرگرفته شده و فرکانس رخدادشان مجزا محاسبه میشود که مسلما اشتباه است.

مشکل دیگر این روشها هم‌آوایی ها هستند. هم‌آوایی ها به کلماتی اطلاق می شود که در جملات مختلف معنایی مختلفی دارند. مثلا کلمه "شیر" بسته به جمله ای که در آن بکار رفته است می تواند به معانی "شیر نوشیدنی"، "شیر جنگل" و "شیر آب" بکار رود. در روشهای آماری تمامی این معنایی یکسان در نظر گرفته میشود.مشکل دیگر این روشها خوانایی پایین و عدم پیوستگی مطالب میباشد. چون تمامی این روشها مبتنی بر تکنیک‌های آماری بوده و توجهی به ارتباط معنایی بین جملات ندارند.

۵-۲ روش های خلاصه سازی مبتنی بر تکنیک های هوش مصنوعی

با توجه به اهمیت موضوع خلاصه سازی متن، تقریبا اکثر تکنیک‌های هوش مصنوعی در خلاصه سازی بکار برده شده است.در مقالات بسیار زیادی شاهد استفاده از روشهای فازی برای خلاصه سازی هستیم.
[Azzam,1999][Dong,1979][Graesser,1981][Mckeown,1995] [Schilder,1977][Yong,1985] شبکه‌های عصبی هم در مقالات زیادی استفاده شده است.

۵-۲-۱ شبکه‌های عصبی مصنوعی: این روش‌ها که دیگر قدمت نسبتاً زیادی نیز دارند،[Svore,2007][kaikhah,2004] با برخورداری از پایه‌های مستحکم در علم ریاضی و مهندسی کنترل، سعی می‌کنند با ساختاری الهام گرفته از ساختار شبکه عصبی مغز انسان به ارتباط معنادار میان ورودی‌ها و خروجی‌های مطلوب برای داده‌ها دست بیابند. به طور کلی این روش‌ها با تنظیم پارامترهای مختلفی نظیر وزن میان نرون‌ها و ثوابت عددی تاثیرگذار در توابع فعال سازی، نگاشت یاد شده را به صورت تدریجی به گونه‌ای بدست می آورند که اختلاف میان خروجی حاصل از شبکه و خروجی واقعی هر داده (آموزشی) به حداقل ممکن برسد.

۵-۲-۲ الگوریتم ژنتیک: این روش هم در خلاصه سازی زیاد استفاده شده است. [Qazvinian,2008][silla,2004]طول کروموزم-های الگوریتم ژنتیک به تعداد جملات موجود در خلاصه است. ژن‌های هر کروموزم در جمعیت الگوریتم ژنتیک بیانگر شماره جمله‌هایی است که در خلاصه حضور دارند. این سیستم برای ریشه‌یابی اسامی و افعال از ریشه‌یابی کراوتز استفاده می‌کند و قادر به تشخیص کلمات مترادف است.البته ترکیب این روشها با همدیگر را هم در مقالات زیادی میتوان مشاهده کرد.

۵-۳ روشهای خلاصه سازی مبتنی بر روشهای معنایی سطوح بالاتر

این دسته از روشها برای غلبه بر مشکلات روشهای پیشین معرفی شدند و سعی دارند تا با فهم عمیق تر متن و استخراج روابط معنایی پنهان موجود در متن، جملات مهم را با دقت بیشتری انتخاب نمایند. از جمله این روشها می توان به روش مبتنی بر ساختارهای معنایی/ نحوی سطح بالا ، روشهای مبتنی بر شبکه یا گراف و روشهای مبتنی بر آنالیز روابط معنایی پنهان موجود در متن اشاره نمود. در ادامه به هریک از این دسته از روشها اشاره شده است .

۵-۳-۱ روش مبتنی بر ساختارها ی معنایی/ نحو ی سطح بالا

ادعای این دسته از روشها این است که جملات یا پاراگرافهای مهم، جمله ها و پاراگرافهایی هستند که در ساختار معنایی متن، ارتباطات تنگاتنگی با یکدیگر دارند. این روش را می توان به دسته های زیر تقسیم بندی نمود:

- روش مبتنی بر شباهت لغوی (زنجیره های لغوی ، مبتنی بر شباهت شبکه واژگان)[Barzilay,1997]
- روش مبتنی بر کلمات هم ارجاع [Azzam,1999]
- روش مبتنی بر استخراج کلمات هم رخداد[Zamanifar,2008]

روش مبتنی بر شباهت لغوی : شباهت معنایی بین کلمات را برروی جملات بسط داده و شباهت بین جملات را محاسبه می شود. یعنی برای محاسبه شباهت بین دو جمله مثلا میتوان شباهت معنایی مبتنی بر شبکه واژگان بین تمامی کلمات دو جمله را محاسبه کرد و پس از محاسبه شباهت بین جملات به صورت ترکیبی از سایر روشها برای انتخاب بهترین جمله ها استفاده نمود. روش زنجیره های لغوی از جمله روشهای معروف در خلاصه‌سازی متن می باشد [Barzilay,1997] برای توضیح بیشتر این روش در ابتدا تعاریفی شرح داده می شود :
پیوستگی لغوی (تعریف۱) پیوستگی لغوی را می توان در قالب شباهت مبتنی بر شبکه واژگان تعریف کرد. پیوستگی لغوی میزان پیوستگی و ارتباط بین دو کلمه را مشخص میکند و شامل دو دسته اصلی می شود:

۱). تصریحی: که شامل مواردی نظیر هم خانواده ها ، متضادها ،شمول و ... میباشد.

۲). هم رخدادی ها: که شامل کلماتی است که با هم در متن رخ میدهند نظیر کلمات "معلم" و "مدرسه" در جمله "او به عنوان معلم در مدرسه کار می کند ".

زنجیره لغوی (تعریف۲): زنجیره لغوی به توالی‌ای از کلمات اطلاق میگردد که با یکدیگر پیوستگی لغوی دارند .برای خلاصه سازی با استفاده از زنجیرههای لغوی گامهای زیر انجام میشود:

۱. در ابتدا کلمات کاندید انتخاب میشود. عموما در مقالات اسامی را به عنوان کلمات کاندید در نظر می گیرند[barzilay,1997] . بنابراین در ابتدا در فاز پیش پردازش از عملیات برچسب زنی بخش‌های سخن استفاده میشود تا نوع کلمات مشخص شود. سپس اسامی به عنوان مجموعه کلمات کاندید در نظر گرفته میشوند.

۲. برای هر یک از کلمات کاندید زنجیره مناسب بر اساس میزان ارتباط بین کلمات زنجیر و کلمه کاندید یافته میشود.

۳. اگر چنین زنجیری پیدا شد، کلمه کاندید در آن قرار میگیرد، در غیر اینصورت کلمه در یک زنجیر جدید قرار میگیرد.

۴. سپس با استفاده از روشهایی تعریف شده قویترین زنجیرها استخراج میگردد.
۵. در ادامه برای تولید خلاصه از راه حل های مختلفی میتوان استفاده نمود. به سه راه حل در زیر اشاره شده است.
۵-۱ یک راه حل این است که اولین جمله ای که یکی از کلمات زنجیرههای قوی را دارد به عنوان نماینده آن زنجیر انتخاب شود.
۵-۲ راه حل دوم برای تولید خلاصه این است که اولین جمله ای را انتخاب نموده که شامل کلمه نماینده زنجیر (مثلا کلمه ای که دارای فرکانس بالایی هست) باشد.
۵-۳ راه حل سوم تعیین قسمت هایی از متن میباشد که چگالی زنجیره (نسبت کلمات زنجیره در متن) در آنجا زیادتر میباشد.در شکل۲، یک نمونه متن و زنجیر های تشخیص داده شده آورده شده است.

> **Mr. Kenny** is the**person** that invented the anesthetic **machine** which uses **micro-computers** to control the rate at which an anesthetic is pumped into the blood. Such **machines** are nothing new. But his **device** uses two **micro-computers** to achineve much closer monitoring of the **pump** feeding the anesthetic into the patient.

شکل ۲- نمونه هایی از زنجیره های لغوی

همانطور که در شکل مشخص است " person " و "mr.kenny " در یک زنجیر و " machain "، "micro-computer"،"device"،"machaines"،"computer"،"pump" در یک زنجیر قرار گرفته اند

روش مبتنی بر کلمات هم ارجاع :

در این مدل خلاصه سازی سعی میشود تا بهترین زنجیره از میان زنجیرههای هم ارجاع برای نمایش موضوع اصلی انتخاب گردد. بنابراین در این روش در گام نخست زنجیره-های هم ارجاع استخراج شده و در گام دوم بهترین زنجیره انتخاب میشود[Azzam,1999] .عبارت های هم ارجاع به عباراتی گفته می شود که علیرغم تفاوت نحوی، به یک نهاد واحد اشاره میکنند. به عنوان مثال عبارات "وزیر امور خارجه آمریکا" و "هیلاری کلینتون" هر دو به یک مرجع اشاره میکنند. ابزارهای مختلفی برای تشخیص این هم ارجاعی ها تاکنون نوشته شده است.
برای انتخاب زنجیره های هم ارجاع هم معیارهای مختلفی ارائه شده است که از جمله این معیارها می توان به موارد زیر اشاره کرد:

- طول زنجیر: این معیار باعث میشود که زنجیره هایی که بیشترین ورودی را دارند انتخاب شوند. با اعمال این معیار نمونه هایی که در متن زیاد تکرار شده اند انتخاب میشوند . در مواردی که طول زنجیرها یکسان باشد از معیار وسعت زنجیر استفاده میشود.
- وسعت زنجیر: این معیار فاصله بین اولین و آخرین عضو زنجیر را محاسبه میکند. زنجیره ایی که وسعت بیشتری از متن را پوشش داده باشد به عنوان زنجیره برتر انتخاب میگردد.
- شروع زنجیر: این معیار زنجیره ای را انتخاب میکند که به واحدی در عنوان متن یا ابتدای متن اشاره کند.

این روش برای خلاصه سازی تک سندی مناسب میباشد اما در خلاصه سازی چندسندی به دلیل وجود اسناد مختلف در ورودی ، مشکلاتی دارد.دشوار بودن استخراج زنجیره ها یا عدم وجود زنجیره ها و همچنین مناسب نبودن بعضی از ویژگی های انتخاب زنجیره از جمله این مشکلات میباشد.

روش مبتنی بر استخراج کلمات هم رخداد:

این دسته از روشها مبتنی بر استخراج کلمات هم رخداد می- باشند . کلمات هم رخداد کلماتی هستند که در متن در قسمتهای مختلف معمولا با یکدیگر ظاهر میشوند و با همدیگر ارتباط معنایی دارند مثل کلمات "سیب" و "چاقو" . معیارها و روشهای مختلفی برای محاسبه هم رخدادیها وجود دارد. درجه هم رخدادی بین دو کلمه از طریق رابطه (۳)محاسبه می شود:

$$R(w_x, w_y) = \frac{f(w_x, w_y)}{f(w_y)} \qquad (۳)$$

که f (Wx,Wy) تعداد دفعاتی است که کلمات Wx و Wy با هم در یک واحد معنایی (به عنوان مثال جمله یا پاراگراف) ظاهر می شوند.پس از استخراج روابط بین کلمات، غالبا گراف مربوط به کلمات و روابطشان ساخته میشود. سپس با روشهای مختلف ،مهمترین دسته از کلمات

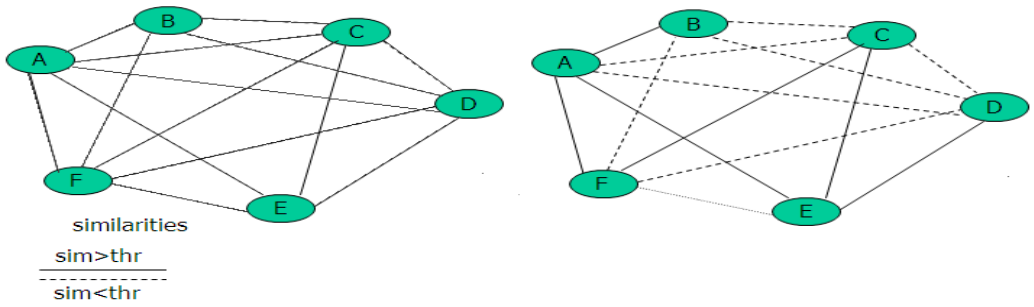
استخراج میگردد. جملاتی که حاوی کلمات مهم باشند به عنوان خلاصه انتخاب میگردند. یکی از مشکلا ت عمده این مدل پیچیدگی محاسبات میباشد. پیدا کردن بهترین روابط بین چندین کلمه در گراف ساخته شده هم از مباحث مهم میباشد.

۶ –روشهای مبتنی بر شبکه ای گراف

در این دسته از روشهای خلاصه سازی، واحدهای متن نظیر جمله و یا پاراگراف به صورت برداری از کلمات وزن دار $D_i = (d_{i1}, d_{i2}, \dots, d_{in})$ نمایش داده میشوند. سپس شباهت بین جملات با استفاده از روش‌های اندازه‌گیری شباهت برداری نظیر ضرب داخلی یا فاصله ی کسینوسی محاسبه میگردد:

$$sim(D_i, D_j) = \sum d_{ik} . d_{jk} \qquad (۴)$$

برای ساخت گراف وزن‌دار، واحدهای متن نظیر جمله یا پاراگراف به عنوان گره‌های گراف در نظر گرفته می شود. برای وزن دهی به یالهای بین دو گره i,j از میزان شباهت بین دو گره یا همان sim(di,dj) استفاده می شود. پس از ساخت گراف اولیه، یال هایی که اهمیت کمتری دارند حذف می شوند. برای همین منظور یالهایی که وزنشان از مقدار حد آستانه α (که توسط الگوریتم تعیین می شود) کمتر باشند حذف شده و گراف نهایی ساخته میشود.



شکل ۳-نمایش گراف جملات

در ادامه ی پردازش، خلاصه از روی گراف نهایی ساخته میشود. برای ساخت خلاصه از روی گراف هم راهکارهای مختلفی پیشنهاد شده است که در ذیر به چند نمونه از آنها اشاره شده است :

- روش انبوه ترین مسیر: در این روش برای انتخاب جملات مهم از گره ای شروع میشود که بیشترین ارتباط یا لینک را با سایر گره ها دارد. سپس گره دومی را که ارتباط بیشتری با سایر گرهها دارد به عنوان جمله یا پاراگراف دوم به خلاصه اضافه میشود. این عملیات آنقدر ادامه پیدا میکند که میزان خلاصه مورد نظر تولید شود. سپس جملات بر اساس موقعیتشان در متن دوباره مرتب میشوند. در حقیقت ایده اصلی این روش این است که واحدهای مهم و برجسته ، بخشهایی از متن هستند که باسایر قسمت ها ارتباطات زیادی دارند.
- روش اول عمق : در این مدل سعی شده است تا جملات خلاصه ارتباطات بالایی داشته باشند و علاوه بر آن خلاصه انتخاب شده دارای پیوستگی مناسبی هم باشد. به همین دلیل برای انتخاب بهترین جمله ها، در ابتدا گره ای که بیشترین لینک را دارد انتخاب میشود. سپس از بین گره هایی که به گره انتخاب شده متصل هستند، گره های که بیشترین لینک را دارد به عنوان جمله دوم انتخاب می‌شود و به همین ترتیب الگوریتم به صورت اول عمق جلو میرود. به عنوان مثال در شکل ۳، ابتدا گره(جمله) A انتخاب میشود چون دارای ۳ لینک بوده و بیشترین لینک را دارد. سپس از بین گرههایی که به A متصل هستند یعنی گرههای B،E و F گرهایی که بیشترین لینک را دارد (گرهE) انتخاب میشود و به همین ترتیب عملیات دنبال میشود.

با توجه به ویژگیهای روشهای مبتنی بر گراف، مسلما این دسته از روشهای خلاصه سازی را به راحتی می توان با سایر روشهای خلاصه سازی ادغام کرده و نتایج مناسبی حاصل نمود. مثلا بجای نمایش برداری جملات و استفاده از فاصله کسینوسی برای وزن دهی به یالها، میتوان از

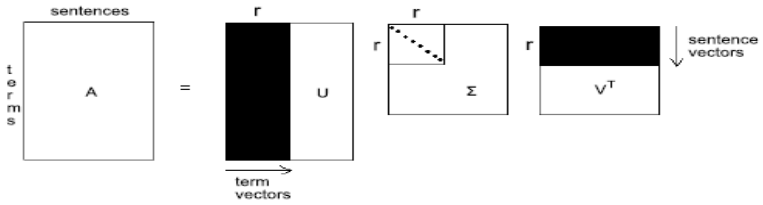
شباهتهای مبتنی بر زنجیره لغوی استفاده نمود و سایر عملیات را دنبال کرد. مقالات زیادی منتشر شده اند که در آنها از روشهای گرافی به صورت ترکیبی با سایر روشها استفاده شده است.[Erkan,2004][Mihalcea,2006,2005]

۷- روش های مبتنی بر آنالیز روابط معنایی پنهان در متن(LSA)

مقالات زیادی [Gong,2001][Yeh,2005[Steinberger,2004,2005]][Young,1985] برای خلاصه سازی از این روش استفاده کرده اند. روش آنالیز روابط معنایی پنهان برای حل مشکل تنک بودن داده ها ارائه شده است. این روش با نمایش داده ها در فضای معنایی کوچکتر و در حقیقت کاهش ابعاد، تا حد زیادی مشکل داده‌های تنک را حل می کند. این کاهش ابعاد منجر به بهبود کارایی در بسیاری از کاربردها شده است . [Deerwester,1990][Papadimitrious,2000]

در تمامی روشهایی که از LSA برای خلاصه سازی استفاده می کنند بردار جملات را بااستفاده از اعمال SVD برروی ماتریس کلمه -جمله میسازند. روش LSA جملات را مستقیماً با استفاده از این ماتریس و با توجه به ویژگیهای معنایی آنها مرتب و دسته بندی می‌کند. lsa در ابتدا برای حل مشکل هم خانواده ها و هم آوایی‌ها در بازیابی اطلاعات معرفی شد.از آن پس LSA در کانون توجه محققان در زمینه پردازش متن قرار گرفت و در بسیاری از مقالات به صورت تئوری و عملی آنالیز گردید . روشLSA در درون خود از روشsvd استفاده می کند.

Svd ماتریس A با ابعاد m*n را به سه ماتریس A=USV تجزیه می کند که U=[Ui,j] ماتریس متعامد ستونی m*m بوده وستون های آن بردار یکه سمت چپ نامیده می شوند. $S = diagonal(\sigma_1, \sigma_2, \ldots, \sigma_n)$ ماتریس قطری n*n است عناصر قطر اصلی آن مقادیر یکه غیر منفی می باشند که به صورت نزولی مرتبط شده اند و V=[Vi,j] هم ماتریس متعامد ستونی بوده و ستون های آن بردارهای یکه سمت راست نامیده می شوند.



اگر r= rank(A)باشد آنگاه حالت روبه رو برقرار خواهد بود: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r \geq \sigma_{r+1} = \cdots = \sigma_n = 0$ (۵)

از دیدگاه پردازش زبان طبیعی ،SVD باعث استخراج روابط معنایی پنهان در ساختار داده می شود.از LSA در تعدادی از روش های خلاصه سازی متن هم استفاده شده است که در ادامه بـه آنهـا اشـاره شـده است:

۱.) اولین بار[Gong,2001] از LSA در خلاصه سازی استفاده نمود وی در ابتدا ماتریس کلمه- جمله را تشکیل داد. برای اینکار معیارهای مختلف وزن دهی را مورد بررسی قرار داد و نهایتا از معیار وزن دهی باینری استفاده نمود . در این معیار اگر یک کلمه در جمله حضور داشته باشد وزن یک و در غیر این صورت وزن صفر میگیرد.پس از تولید ماتریس کلمه-جمله، روش SVD برروی ماتریس اعمال میشود. برای انتخاب جملات مهم از ماتریس بردار های یک سمت راست(V^Tماتریس) استفاده می شود. ایده اصلی روش Gong این است که با اعمال SVD برروی ماتریس کلمه-لغت، موضوعات یا زیر موضوعات پنهان موجود در متن توسط ماتریس V^t (که آن را می توان ماتریس جمله-موضوع نامید) استخراج میشود. بر همین اساس برای انتخاب جمله مهم k ام، ستونی که در سطر k ام ماتریسV^T بیشترین مقدار را داشته باشد به عنوان نماینده k امین موضوع پنهان موجود در متن انتخاب میشود این عملیات تا جایی که به حجم خلاصه مورد نظر برسد ادامه می یابد.این روش دارای ایراداتی است که در مقالات مختلف به آن اشاره شده است. یکی از این ایرادات این است که بین موضوعات مختلف ارزش یکسانی قائل میشود و برای هر کدام یک جمله را به عنوان نماینده انتخاب می کند. بعبارت دیگر برای تمام موضوعات ارزش یکسانی قائل هست و به اندازه مقادیریکه توجهی نمیکند. ایراد دومی که به این روش وارد میشود این است که جمله ای که دارای مقدار بالایی در بردارهای ماتریس V^Tهست ولی بزرگترین مقادیر را ندارد (در هیچ یک از ابعاد شانس برنده شدن ندارد) هرگز برای خلاصه انتخاب نمیشود. برای حل این دو مشکل روشی در [Steinberger,2004] ارایه شد.اما مشکل مهم دیگراین است که مبنا قرار دادن جمله به عنوان عنصر محوری، باعث میشود که بردارهای استخراج شده از خروجی SVD قادر به بیان مفاهیم اصلی پنهان از دیدگاه کلی نباشد چراکه در این روش ها فرکانس کلمات در سطح جمله مورد بررسی قرار میگیرد و تاثیر جملات بر یکدیگر در سطوح بالاتر لحاظ نمیشود.

۲.) در مقاله ی [Steinberger,2004] برای حل دو مشکل اشاره شده برای روش [Gong,2001]، روشی ارائه شد که در آن از اثر مقادیری که برای برجسته کردن اهمیت ابعاد (یا همان موضوعات پنهان) استفاده شده است. در مقاله [ding,2005]نشان داده شده است که اهمیت آماری هر یک از ابعاد خروجی LSA تقریبا متناسب با مجذور مقادیری که متناسب با آنها در ماتریس S می باشد. بر همین اساس در تاثیر

مقادیری که در انتخاب جملات با ضرب مجذور ماتریس مقادیریکه در ماتریس V^Tمحاسبه می شود. بر همین اساس در [Steinberger,2004] تاثیر مقادیر یکه در انتخاب جملات با ضرب مجذور ماتریس یکه در ماتریس V^T و به صورت زیر محاسبه می شود

$$B = \begin{pmatrix} v_{1,1}\sigma_1^2 & v_{1,2}\sigma_1^2 & \dots & v_{1,n}\sigma_1^2 \\ v_{2,1}\sigma_2^2 & v_{2,2}\sigma_2^2 & \dots & v_{2,n}\sigma_2^2 \\ \dots & \dots & \dots & \dots \\ v_{r,1}\sigma_r^2 & v_{r,2}\sigma_r^2 & \dots & v_{r,n}\sigma_r^2 \end{pmatrix} \quad (7) \qquad B = S^2 . V^T \quad (6)$$

به جای انتخاب جملات با بیشترین مقدار در سطرهای ماتریس V^T ، جملاتی که وزن ترکیبی آنها در بین تمامی موضوعات بیشتر باشد انتخاب می شوند. بر همین اساس طول جمله k ام به صورت زیر محاسبه می شود

$$S_k = \sqrt{\sum_{i=1}^{r} b_{i,k}^2} \quad (8)$$

در نهایت جملاتی که بیشترین طول را داشته باشند به عنوان جملات مهم بر گردانده می شوند. میزان r هم برابر r = p / 100*n در نظر گرفته شده است که p، درصد خلاصه سازی میباشد. اگرچه این روش تا حدی مشکل روش [Gong,2001]، را بهبود می بخشد، اما بدلیل عدم توجه به روابط کلی بین جملات، همچنان دارای مشکل میباشد.

۳) در مقاله‌ی دیگری [Steinberger,2004] با دخیل کردن روش ادغام ضمایر در LSA، کارایی روش Gong را افزایش داد. ایده اصلی در آن ادغام اطلاعات لغوی با روش SVD بود. او در این مقاله دو روش جانشینی و روش افزایشی را پیشنهاد کرد. در روش اول از ادغام ضمایر به عنوان یکی از مراحل پیش پردازشی استفاده میشود. یعنی در ابتدا ضمایر با کلمات اصلی جایگزین شده و سپس ماتریس ورودی SVD برروی متن اصلاح شده ساخته می شود. البته نتایج بررسی های منتشر شده بیانگر این است که این روش نه تنها موجب افزایش دقت چندانی در کارایی نمی شود، بلکه در مواردی هم باعث کاهش دقت میشود. در روش دوم یا همان روش افزایشی ، به ماتریس ورودی SVD یک بخش دیگر که شامل اطلاعات لغوی مربوط به زنجیره های متوالی می باشد هم اضافه می شود. وی در این مقاله ثابت کرد که ادغام ضمایر با استفاده از روش افزایشی، دقت خلاصه سازی روش Gong را افزایش میدهد. لازم به ذکر است که این روش راهکاری در جهت حل مشکلات اصلی ذکر شده در قسمت قبل ارائه نمی کند.

۴) در [Yeh,2005] از ترکیب نگاشت رابطه ای متن و آنالیز روابط معنایی پنهان در متن برای تولید خلاصه استفاده شد. وی با استفاده توامان از LSA و T.R.M(نگاشت رابطه ای متن) ساختار معنایی پنهان موجود در سند را استخراج کرده و خلاصه را بر اساس این ساختار پنهانی استخراج شده تولید نمود. این ترکیب منجر به افزایش دقت در تولید خلاصه شده است.

۵) در [young,1985] روشی برای خلاصه سازی اخبار مبتنی بر LSA ارائه شد. وی در ابتدا با استفاده از lsa شباهت معنایی بین جملات را محاسبه کرده و سپس با استفاده از روش خوشه بندی مبتنی بر روش k-means جملات را دسته بندی کرده و سپس از هر کلاستر، جمله ای که بیشترین شباهت را به موضوع داشته باشد به عنوان نماینده آن خوشه بر می گرداند.که این موضوع تا حدی مشکل Gong را بر طرف می کند.

نتیجه گیری

این تحقیق روش های خلاصه کردن را ارائه می دهد.یک خلاصه از نوع استخراجی جملات مهم را از متن اصلی انتخاب می کند.این جملات با اهمیت با استفاده از ویژگی های زبانی و آماری استخراج می شوند.مشکل روش های خلاصه سازی استخراجی این است که در بسیاری از موارد ما شاهد افزونگی یا همان جملات تکراری در متن استخراج شده هستیم.در این موارد در صورتی که برخی جملات را حذف شوند ،میزانی از اطلاعات را از دست می دهیم.ودر صورتیکه همه جملات آورده شوند ،دچار افزونگی اطلاعات خواهیم شد و از هدف خلاصه سازی دور می شویم.نرم افزار های خلاصه ساز متن یک خلاصه تاثیر گذار تولید می کنند که منجر به استفاده از منابع بیشتربا سرعت بالاتر ودر نتیجه حاصل شدن اطلاعات غنی تر می شود.ارزیابی سیستم های خلاصه ساز متن به طور کلی به دو روش بیرونی و درونی تقسیم بندی می شوند.روش ارزیابی بیرونی بیشتر بر استفاده از خلاصه ها در کاربردهای خاص نظیر اجرای دستورات ،بازیابی اطلاعات،پاسخ دهی به سوالات و ارزیابی ارتباط بین خلاصه وسند اصلی متمرکز می باشند.در روش ارزیابی درونی تمرکز بیشتر بر روی پیوستگی مطالب و میزان اطلاع رسانی آن ها می باشد و معمولا شامل مقایسه کیفیت خلاصه تولید شده توسط سیستم و خلاصه های انسانی مب باشد.روش های ارزیابی درونی انسانی ،کیفیت اسناد خلاصه با ارزیابی مغیار های دقت،خوانایی و وضوح تعیین می نمایند.

منابع

[1] اخوان تارا ،شمس فرد مهرنوش،عرفانی مونا،Parsamist خلاصه سازتک سندی و چند سندی متون فارسی ،چهاردهمین کنفرانس انجمن کامپیوتر ایران،تهران،اسفند ۱۳۸۸

[2] پور غلامی فاطمه ،کاهانی محسن،پور معصومی آصف، خلاصه سازی چکیده ای مبتنی بر مشابهت جملات،نخستین کنفرانس بین المللی پردازش خط زبان فارسی شهریور ۱۳۹۱

[3] پور معصومی آصف، کاهانی محسن، کامیار محسن،کامیار حسین ،خلاصه سازی خودکار چند سندی مبتنی بر استخراج مفاهیم،انجمن کامپیوتر ایران،۱۳۹۱

[1] Karel Jezek and Josef Steinberger, "Automatic Text summarization", Vaclav Snasel (Ed.): Znalosti 2008, pp.112, ISBN 978-80-227-2827-0, FIIT STU Brarislava, Ustav Informatiky a softveroveho inzinierstva, 2008

[2] Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami and Pooya Khosravyan Dehkordy, "Optimizing Text Summarization Based on Fuzzy Logic", In proceedings of Seventh IEEE/ACIS International Conference on Computer and Information Science, IEEE, University of Shahid Bahonar Kerman, UK, 347-352, 2008

[3] Joel larocca Neto, Alex A. Freitas and Celso A.A.Kaestner, "Automatic Text Summarization using a Machine Learning Approach", Book: Advances in Artificial Intelligence: Lecture Notes in computer science, Springer Berlin / Heidelberg, Vol 2507/2002, 205-215,2002

[4] Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, "Tapping into the Power of TextMining",Journal of ACM, Blacksburg, 2005.

[5] Fang Chen, Kesong Han and Guilin Chen, "An Approach to sentence selection based text summarization", Proceedings of IEEE TENCON02, 489-493, 2002.

[6] Mohamed Abdel Fattah and Fuji Ren, "Automatic Text Summarization", Proceedings of World Academy of Science, Engineering and Technology, Vol 27,ISSN 1307-6884, 192-195, Feb 2008.

[7] H. P. Luhn**, "**The Automatic Creation of Literature Abstracts", Presented at IRE National Convention, New York, 159-165, 1958.

[8] H. P. Edmundson.," New methods in automatic extracting", Journal of the ACM, 16(2):264-285, April 1969.

[9] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer", In Proceedings of the 18th ACMSIGIR Conference, pages 68-73, 1995

[10] Ronald Brandow, Karl Mitze, and Lisa F. Rau."Automatic condensation of electronic publications by sentence selection. Information Processing and Management", 31(5):675-685,1995.

[11] E. Mittendorf and P. Schauble, " Document and passage retrieval based on hidden markov models", In Proceedings of the 17th ACM-SIGIR Conference, pages 318-327,1994.

[12] A. Bookstein, S. T. Klein, and T. Raita, "Detecting content-bearing words by serial clustering", In Proceedings of the 18th ACM-SIGIR Conference, pages 319-327, 1995.

[13] Madhavi K. Ganapathiraju, "Overview of summarization methods", 11-742: Self-paced lab in Information Retrieval, November 26, 2002.

[14] Klaus Zechner, "A Literature Survey on Information Extraction and Text Summarization", Computational Linguistics Program, Carnegie Mellon University, April 14, 1997.

[15] Chin-Yew Lin and Eduard Hovy," From Single to Multidocument Summarization: A Prototype System and its Evaluation", Proceedings of the ACL conference, pp. 457–464. Philadelphia, PA. 2002.

[16] Rene Arnulfo Garcia-Herandez and Yulia Ledeneva, "Word Sequence Models for Single Text Summarization", IEEE,44-48, 2009.

[17] Yongzheng, Nur and Evangelos, "Narrative Text Classification for Automatic Key Phrase Extraction in Web Document Corpora", WIDM'5, 51-57, Bremen Germany,2005.

[18] Canasai Kruengkari and Chuleerat Jaruskulchai, "Generic Text Summarization Using Local and Global Properties of Sentences", Proceedings of the IEEE/WIC international Conference on Web Intelligence (WI'03) , 2003.

[19] Meng Wang, Xiaorong Wang and Chao Xu, "An Approach to Concept Oriented Text Summarization", in Proceedings of ISCIT'05, IEEE international conference, China,1290-1293, 2005.

[20] Azadeh Zamanifar, Behrouz minaei-Bidgoli and Mohsen Sharifi," A New Hybrid Farsi Text Summarization Technique Based on Term Co-Occurrence and Conceptual Property of Text ", In Proceedings of Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, IEEE, 635-639, Iran, 2008.

[21] Khosrow Kaikhah, "Automatic Text Summarization with Neural Networks", in Proceedings of second international Conference on intelligent systems, IEEE, 40-44, Texas, USA, June 2004.

[22] Ladda Suanmali, Mohammed Salem, Binwahlan and Naomie Salim, "Sentence Features Fusion for Text summarization using Fuzzy Logic, IEEE, 142-145, 2009
[23] David B. Bracewell, Fuji REN and Shingo Kuriowa, "Multilingual Single Document Keyword Extraction for Information Retrieval", Proceedings of NLP-KE'05, IEEE, Tokushima, 2005.
[24] David Kirk Evans, "Identifying Similarity in Text: Multi Lingual Analysis for Summarization", PhD thesis, Graduate School of Arts and Sciences, Columbia University, 2005.
[25] Cowie, J., Mahesh, K., Nirenburg, S., and Zajaz, R., "MINDS-Multilingual INteractive document summarization", In Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization (pp. 131– 132). Menlo Park, CA: AAAI, 1998.
[26] Dragomir Radev, Timothy Allison, Sasha Blair- Goldensohn, John Blitzer, Arda C¸ elebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, and Zhang Zhu., "MEAD - a platform for multi document multilingual text summarization", In Proceedings of LREC 2004, Lisbon, Portugal, May 2004.
[27] Khosrow Kaikhah "Text Summarization using Neural Networks", Department of Faculty Publications- Computer Science, Texas State University, eCommons,2004.
[28] Ladda Suanmali, Naomie Salim and Mohammed Salem Binwahlan, "Fuzzy Logic Based Method for Improving Text Summarization", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 2, No.1,2009
[29] Rasim M. Alguliev and Ramiz M. Aliguliyev, "EffectiveSummarization Method of Text Documents", in Proceedings of IEEE/WIC/ACM international conference on Web Intelligence (WI'05), 1-8, 2005.
[30] Berry Michael W., "Automatic Discovery of Similar Words", in "Survey of Text Mining: Clustering,Classification and Retrieval", Springer Verlag, New York, LLC, 24-43, 2004.
[31] Vishal Gupta, G.Sl Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, VOL. 1, NO. 1, 60-76, AUGUST 2009.
[32] G Erkan and Dragomir R. Radev, "LexRank: Graph-based Centrality as Salience in Text Summarization", Journal of Artificial Intelligence Research, Re-search, Vol. 22, pp. 457-479 2004.
[33] Udo Hahn and Martin Romacker, "The SYNDIKATE text Knowledge base generator", Proceedings of the first International conference on Human language technology research**,** Association for Computational Linguistics , ACM, Morristown, NJ, USA , 2001.
[34] Ani Nenkova and ,Rebecca Passonneau, "Evaluating content selction in summarization: The Pyramid method", in HLT-NAACL, 145-152, 2004.
[35] Kathleen Mackeown, Ani Nenkova, David Elson, Rebecca Passonneau, and Julia Hirschberg "A task based evaluation of multidocument system",in SIGIR'05, ACM,2005
[36] Chin-yew Lin, "A package for automatic evaluation of summaries",in Proc. ACL workshop on text summarization branches out,2004.
[37] Eduard Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto, "Automated Summarization Evaluation with Basic Elements", In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), 2006,
[38] Samuel W K. Chan, Tom B. Y. Lai, W.J. Gao and Benjamin K. T'sou, "Mining discourse structures for Chinese textual summarization**",** NAACL-ANLP workshop on Automatic Summarization, ACM, Seattle, Washington, 11-20, 2000.
[39] Junlin Zhanq, Le Sun and Quan Zhou, "A Cue-based Hub- Authority Approach for Multi-Document Text Summarization", in Proceeding of NLP-KE'05, IEEE,642- 645, 2005
[40] Ben Hachey, "Multi-document summarization using generic relation extraction**",** Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing: Volume 1, 420-429, 2009.
[41] K. S. Jones, "Automatic summarizing: the state of the art," Information Processing and Management, Elsevier, Vol.43, No. 6, pp. 1449–1481, 2007.
[42] F. Canan Pembe and Tunga Güngör, "Automated Querybiased and Structure-preserving Text Summarization on Web Documents", Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications, İstanbul, June 2007.
[43] Ramakrishna Varadarajan and Vagelis Hristidis, "Structure-Based Query-Specific Document Summarization", in proceedings of CIKM'05, ACM, Bremen, Germany, 2005.
[44] Feng Jin, Minlie Huang and Xiaoyan Zhu, " A Queryspecific Opinion Summarization System", in proceedings ofICCI '09, 8th IEEE international conference on cognitive informatics, Kowloon, Hong Kong, 428-433,2009.
[45] Hal Daum´e III and Daniel Marcu, "Bayesian Query- Focused Summarization", Proceedings of the 21st International Conference on Computational Linguisticsand 44th Annual Meeting of the ACL, 305–312, Sydney,July 2006.
[46] Jimmy Lin., "Summarization.", Encyclopedia of Database Systems. Heidelberg, Germany: Springer-Verlag, 2009.

[47] Jackie CK Cheung, "Comparing Abstractive and Extractive Summarization of Evaluative Text: Controversiality and Content Selection", B. Sc. (Hons.) Thesis in the Department of Computer Science of the Faculty of Science, University of British Columbia, 2008.
[48] Conroy, J.M., Schlesinger, J.D., Stewart, J.G. "CLASSYQuery-Based Multi-Document Summarization". In DUC 95 Conference Proceedings,Boston, USA, 2995
[49] Barzilay, R. & Elhadad, M. "Using Lexical Chains for Text Summarization". In Proceedings of the ACL Workshop on Intelligent ScalableText Summarization, pages 19–17, Madrid, Spain,August 1997.
[50] Azzam, S., Humphreys, K., & Gaizauskas, R. "Using coreference chains for text summarization". In Proceedings of the ACL"99workshop on coreference andits applications (pp. 77–84), College Park, MD, USA, 1999.
[51] Erkan, G., Radev, D. R. "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization". Journal of Articial Intelligence Research 22, 457-479 , 2994.
[52] Gong, Y., & Liu, X. "Generic text summarization using relevance measure and latent semantic analysis". In Proceedings of the24th annual international ACM SIGIR conference on research andd evelopment in information retrieval (SIGIR"91) (pp. 19–25), NewOrleans, LA, USA, 2991.
[53] Graesser, A. C. "Prose Comprehension beyondthe Word". NY: Springer-Verlag, 1981.
[54] Kaikhah, K. "Automatic Text Summarization withNeural Networks", in Proceedings of second internationalConference on intelligent systems, IEEE, 49-44, Texas,USA, June 2994.
[55] Kupiec, J., Pedersen, J., Chen, & F. "A trainable document summarizer". In Proceedings of the 18th annual international ACMSIGIR conference on research andd evelopment in information retrieval (SIGIR"95) (pp. 68–73), Seattle, WA, USA, 1995.
[56] Kyoomarsi, F., Khosravi, H., Eslami, E., Khosravyan, P., Tajoddin, A. "Optimizing Text Summarization Based on Fuzzy Logic". Proceedings of the Seventh IEEE/ACIS International Conference on Computer and Information Science , icis, 2998.
[57] Lin, D. "An information-theoretic definition of similarity". In Proceeding of the15th International Conference on Machine Learning, pp. 296–394. Morgan Kaufmann, San Francisco, USA, 1998.
[58] Luhn, H. P. " The automatic creation of literature abstracts". IBM Journal of Research and Development, 2(2), 159–165, 1958.
[59] Mani, I., House, D., Klein, G., Hirshman, L. "The TIPSTER SUMMAC Text Summarization Evaluation". Technical Report MTR 98W9999138, The Mitre Corporation, McLean, Virginia, 1998.
[60] Mihalcea, R. and Tarau, P. "Text-rank - bringing order into texts". In Proceeding of the Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 2004.
[61] Mihalcea, R. and Tarau, P. "An Algorithm for Language Independent Single and Multiple Document Summarization". In Proceedings of the International Joint Conference on Natural Language Processing, Korea, 2005
[62] Mihalcea, R., Corley, C., and Strapparava, C. "Corpus-based and knowledge-based measures of text semantic similarity". In Proceedings of the American Association for Artificial Intelligence.(Boston, MA), 2006.
[63] Schilder, F., Kondadadi, F. "FastSum: Fast and accurate query based multi-document summarization". In Proceedings of the 46th meeting of the Association for Computational Linguistics, Columbus, Ohio, 2008.

[64] Silla, J., Nascimento, C., Pappa, G.L., Freitas, A.A. and Kaestner, C.A.A. „Automatic text summarization with genetic algorithm-based attribute selection", Lecture Notes in Artificial Intelligence, 2004.
[65] Steinberger, J. and Je˘zek, K. "Text Summarization and Singular Value Decomposition". In Proceedings the 3rd International Conference on Advances in Information Systems, Lecture Notes in Computer Science 2457, pp. 245–254, Springer-Verlag, October 2994.
[66] Steinberger, J., & Kabadjov, M.A. & Poesio, M., & Sanchez-Graillet,O . "Improving LSA-based summarization with anaphora resolution". In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. 2005.
[67] Svore, K., Vanderwende, L., and Burges, C. "Enhancing single-document summarization by combining Rank Net and third-party sources". In Proceedings of the EMNLP-CoNLL, 2007.
[68] Dalianis, H. "SweSum A Text Summarizer for Swedish", Technical report", TRITANA-P9915, IPLab-174, 2000.
[69] Yeh, J.Y., Ke, H.R., Yang, W.P., & Meng, I.H. "Text summarization using a trainable summarizer and latent semantic analysis". Information Processing and Management, 41, 75-95, 2005.
[70] Yu, H. "News summarization based on semantic similarity measure". Ninth International Conference on Hybrid Intelligent Systems, vol. 1, pp.180-183, 2000.
[71] Young, S.R., & Hayes, P.J. "Automatic classification and summarization of banking telexes". In Proceedings of the 2ndConference on Artificial Intelligence Applications (pp. 492–498), 1985.